L10 - IP Routing (v6.2)

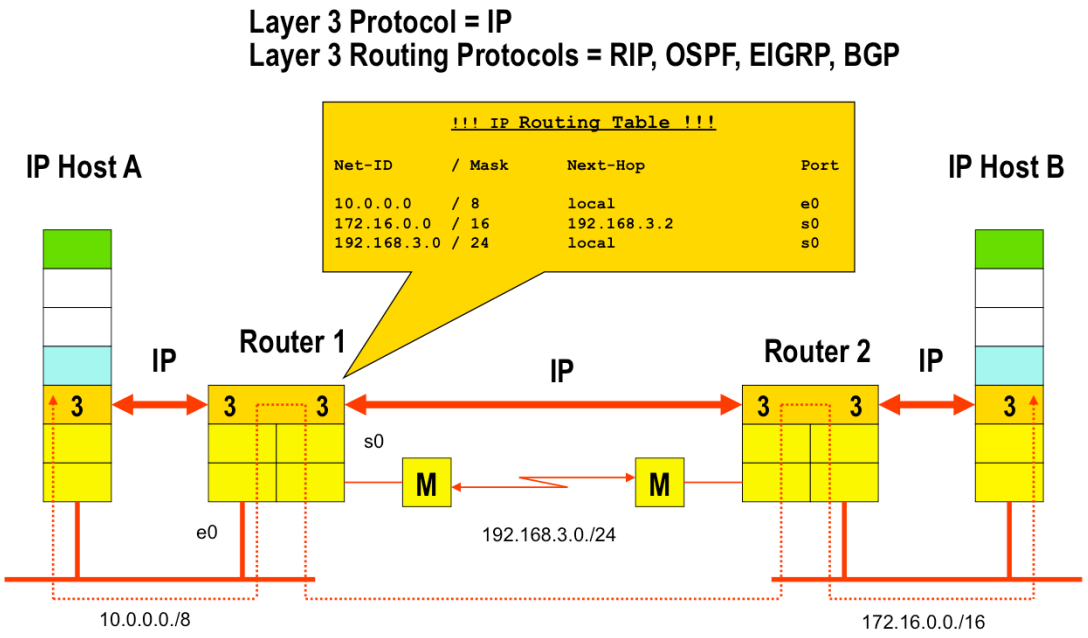
IP Routing

Introduction (Static, Default, Dynamic),
RIP (Distance Vector), OSPF (Link State),
Introduction to Internet Routing (BGP, CIDR)

Agenda

- **Introduction to IP Routing**
  - Basics
  - Static Routing
  - Default Route
  - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

IP, IP Routing Protocol, IP Routing Table



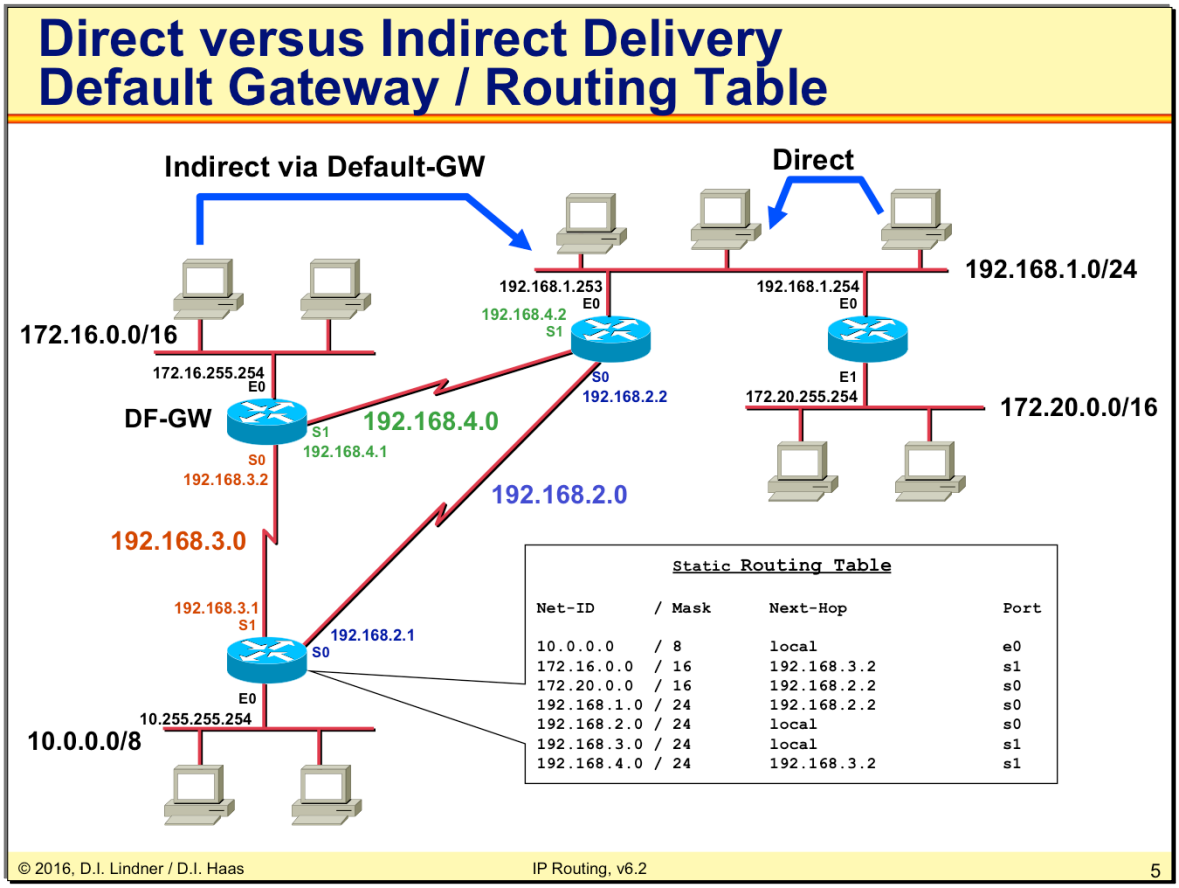
 IP Routing, v6.2 3

What is Routing?

- ***Finding / choosing a path to a destination address***
- **Direct delivery performed by IP host**
  - Destination network = local network
- **Indirect delivery performed by router**
  - Destination network ≠ local network
  - Datagram is forwarded to default gateway
  - Passed on by the router based on routing table
- **Routing table**
  - Database of known destinations
  - Signposts leading to next hop

 IP Routing, v6.2 4

Routing is the process of choosing a path over which to send IP datagrams destined to a given destination address. There are 2 ways to delivery a packet. The direct delivery and the indirect delivery. IP hosts are responsible for direct delivery of IP datagrams whereas routers are responsible for selecting the best path in a meshed network in case of indirect delivery of IP datagrams. IP hosts are further responsible for choosing a default router ("default gateway") as next hop in case of indirect delivery of IP datagrams. When there is a direct delivery (destination network = local network) the host makes for example an ARP-request (Ethernet) and then delivery the datagram to the right host. If there is a indirect delivery (destination network ≠ local network) the IP host forwards the datagram to its default gateway.



Routing table contains signpost as for every known (or specified) destination network:

- net-ID / subnet-mask
- next hop router (and next hop MAC address in case of LAN)
- outgoing port

In the picture above there is small network, and a good example of a routing table. For example a host in network 10 want to send a datagram to a user in network 192.168.1. The destination address ≠ local address so the router must do a forward decision. The router compare the destination address with his routing table and found the right match (192.168.1.0/ 24 192.168.2.2 1 s0). Now he sends out the datagram via port s0 to the next hop, the router with the IP-Address of 192.168.2.2. This router is directly connected to the network 192.168.1.0. After an ARP-request the datagram is delivered to the right user.

IP Routing Paradigm

- **Destination Based Routing**
  - Source address is not taken into account for the forward decision
- **Hop by Hop Routing**
  - IP datagrams follow the path (signpost) given by the current state of routing table entries
- **Least Cost Routing**
  - Typically only the best path is considered for forwarding of IP datagrams
  - Alternate paths will not be used in order to reach a given destination
    - Note: Some methods allow load balancing if paths are equal

 IP Routing, v6.2 6

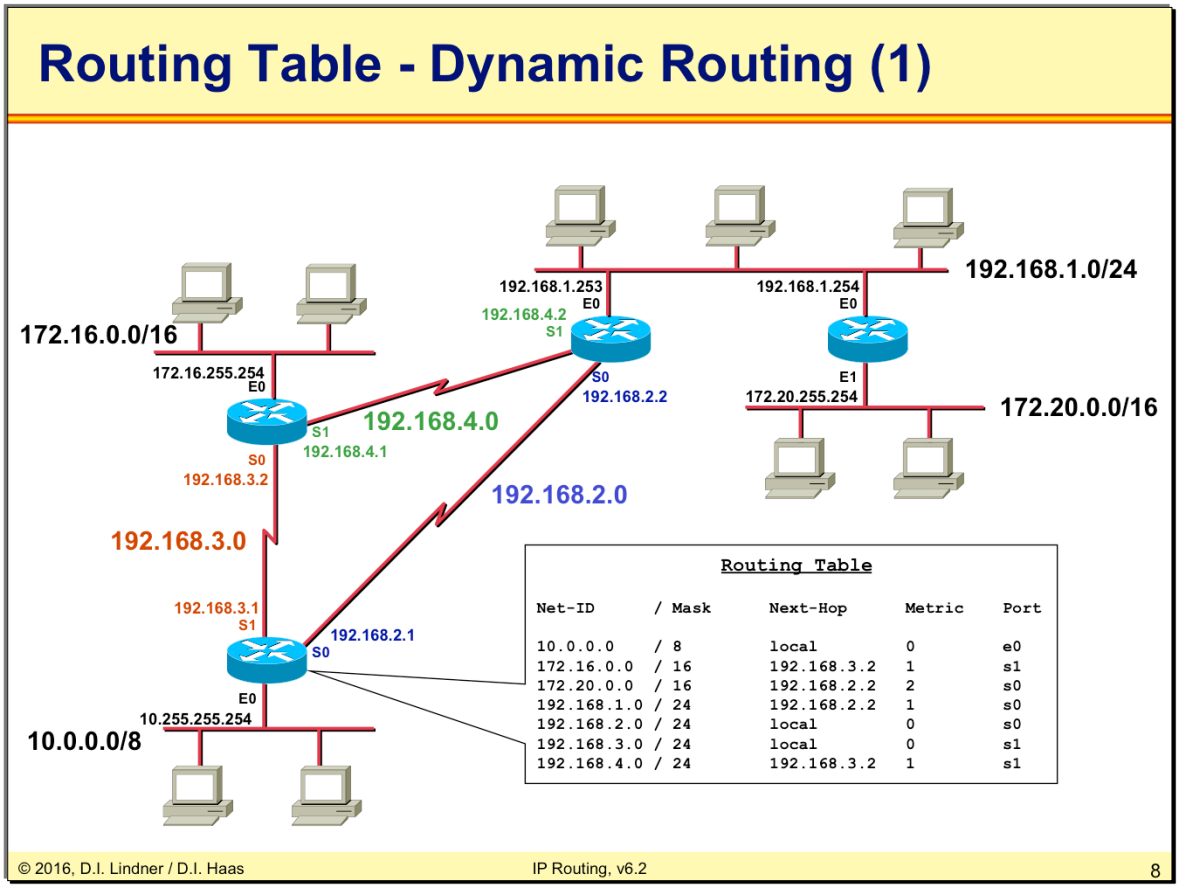
The IP routing paradigm is fundamental in IP routing. Firstly, IP routing is "destination based routing", that means the source IP address is never examined during the routing process. Secondly, IP routing is "hop-by-hop", which emphasizes the difference to virtual circuit principles. The routing table in every router within the autonomous system must be both accurate and up to date (consistent and loop-free) so that datagrams can be directed across the network to their destination.

In IP the path of a packet is not pre-defined and not connection oriented, rather each single router performs a routing decision for each datagram. Thirdly, IP routing is "least cost" in that only that path with the lowest metric is selected in case of multiple redundant paths to the same destination.

Note that several vendors extend these rules by providing additional features, but the routing paradigm generally holds for most of the routers in the Internet, at least for the basic routing processes.

Static versus Dynamic Routing

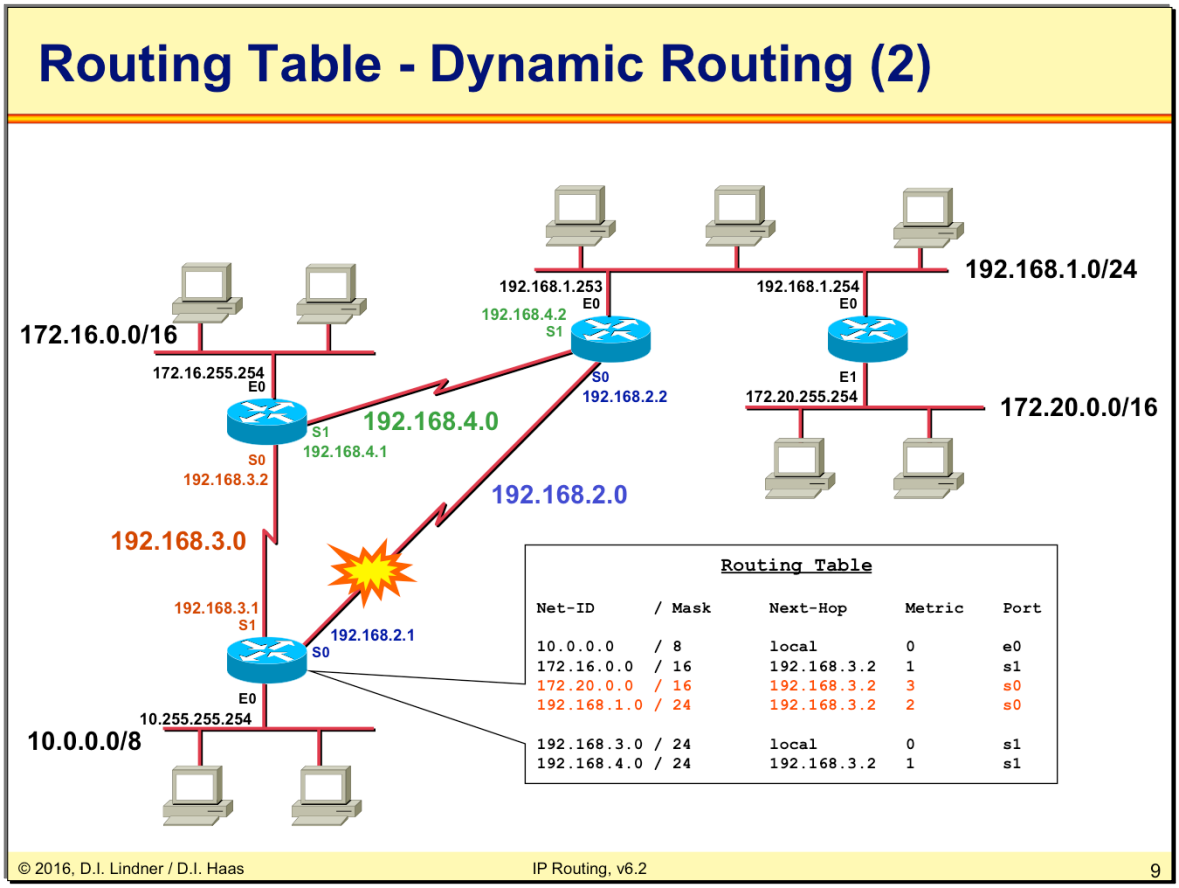
- **Static**
  - Routing tables are preconfigured by network administrator
  - Non-responsive to topology changes
  - Can be labor intensive to set up and modify in complex networks
  - No overhead concerning CPU time and traffic
- **Dynamic**
  - Routing tables are dynamically updated with information received from other routers
  - Responsive to topology changes
  - Low maintenance labor cost
  - Communication between routers is done by routing protocols using routing messages for their communication
  - Routing messages need a certain percentage of bandwidth
  - Dynamic routing need a certain percentage of CPU time of the router
  - That means overhead



Now we see some additional fields in the a routing table bulit by a dynamic routing protocol (in our case RIP with hop counts is assumed):

Routing table contains signpost as for every known (or specified) destination network:

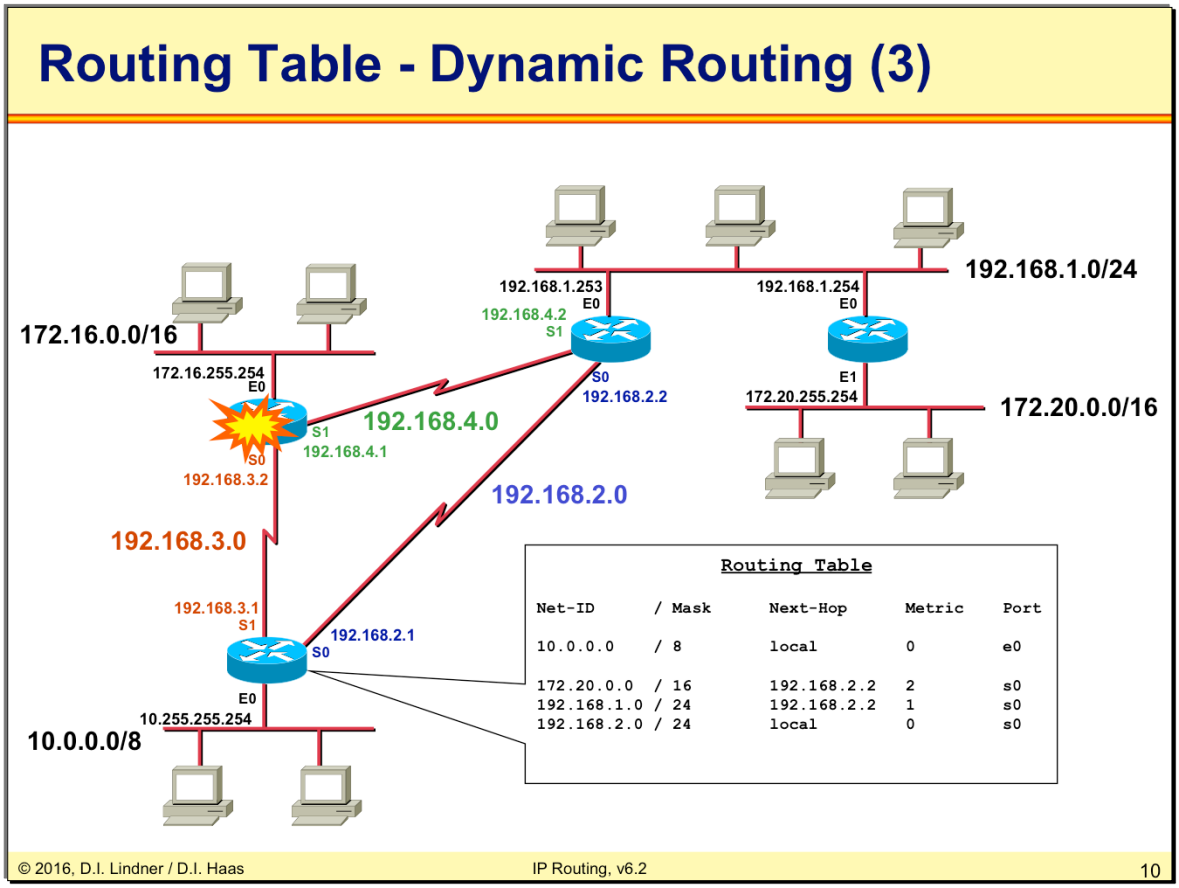
- net-ID / subnet-mask
- next hop router (and next hop MAC address in case of LAN)
- outgoing port
- metric (information how far away is a certain destination network) -> hop counts in our picture
- time reference (information about the age of the table entry)



What can a dynamic routing protocol detect?

Loss of a link between any two directly connected routers
Loss of a router connected in a meshed network
Loss of network directly connected to a router

In our example loss of link 192.168.2.0 causes adaption of the routing table hence traffic from 10.0.0.0 to 192.168.1.0 or 172.20.0.0 will take the alternate = only remaining path via 192.168.3.2. Hop count to these networks has risen by one. If link 192.168.2.0 comes back the dynamic routing will adapt back to picture of last slide.

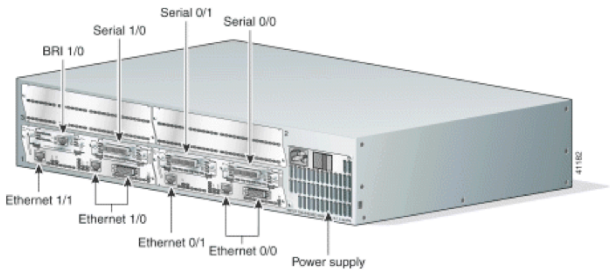


In our example loss of left router causes adaption of the routing table networks 172.16.0.0, 192.168.3.0 and 192.168.4.0 are not longer seen in the routing table If left router comes back the dynamic routing will learn about these network again, hence we can see the automatic appearance of networks in a routing table in case of power on.

IP Router

- Initially Unix workstations with several network interface cards
- Today specialized hardware

Cisco 3600 Router



The picture above shows one of the most used routers today, the Cisco 3600 platform, employing various Ethernet and Serial interfaces.

Agenda

- **Introduction to IP Routing**
  - Basics
  - Static Routing
  - Default Route
  - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Reasons for Static Routing

- **Very low bandwidth links**
- **Link is the only path to a stub network**
- **Dialup links and backup links**
  - X.25 SVC, ISDN, Frame Relay SVC, ATM SVC
- **Administrator needs full control over the link**
  - E.g. for security reasons
  - E.g. in hub and spoke topologies avoiding any-to-any communication
- **Router has very limited resources and cannot run a routing protocol**
- **Cisco syntax:**

ip route *prefix mask* **{***ip-address* **|** *interface-type interface-number***}** **[***distance***]** **[tag** *tag***]** **[permanent]**

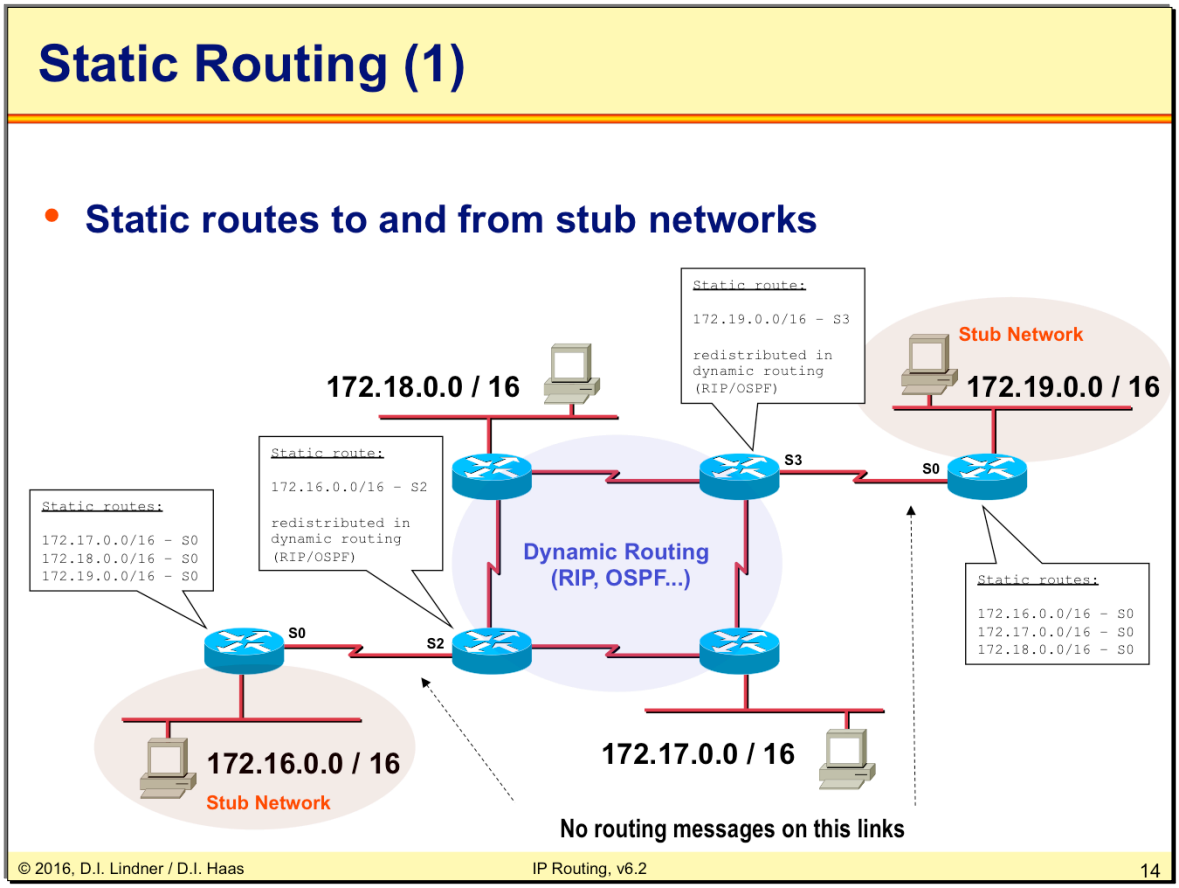
Tag value that can be used as a "match" value for controlling redistribution via route maps

Specifies that the route will not be removed, even if the interface shuts down

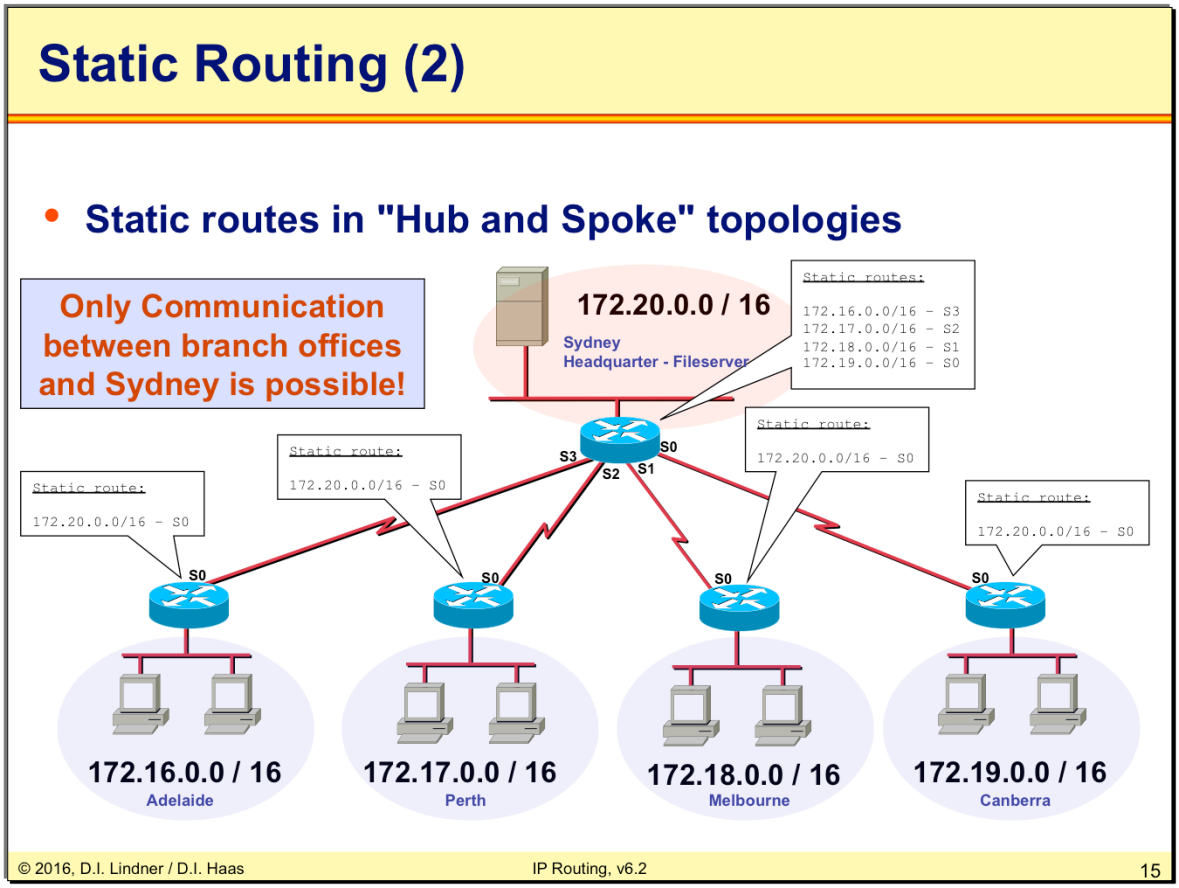
 IP Routing, v6.2 13

For dialup links static routes are the only or preferred way because dynamic routing will keep the link up even if no user traffic exists. The periodic routing messages refreshes the idle-timer which is used for tear down the link after a certain period of silence (no user traffic). The other way would be to open and close the dialup link for routing messages during times of silence. But that would stress the signaling system and you have to pay for every call.

Route redistribution is the method specified by Cisco in order to import routing information into another routing processes (e.g. static routes into the dynamic OSPF process. Route maps can be used to filter routes during the import. Filter maybe based on tags. Distance is the so called administrative distance (AD) which gives a certain level of trust to a route. The smaller the value of AD the better the information about a route. That helps to decide which route to be put into the routing table, if information about the same route is available from different routing processes.



In the picture above we see a good example of static routes. Every router in the stub networks is configured manually, because there is only one way the datagram can go. To allow connectivity to hub sites, the static routes have to be redistributed into the dynamic routing processes (e.g. RIP or OSPF). Otherwise routers without static routes in the above picture will not get information about the stub networks.



Here you see a other example of static routing. Every branch office is connected over static routes with the Sydney headquarter.

Agenda

- Introduction to IP Routing
  - Basics
  - Static Routing
  - Default Route
  - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

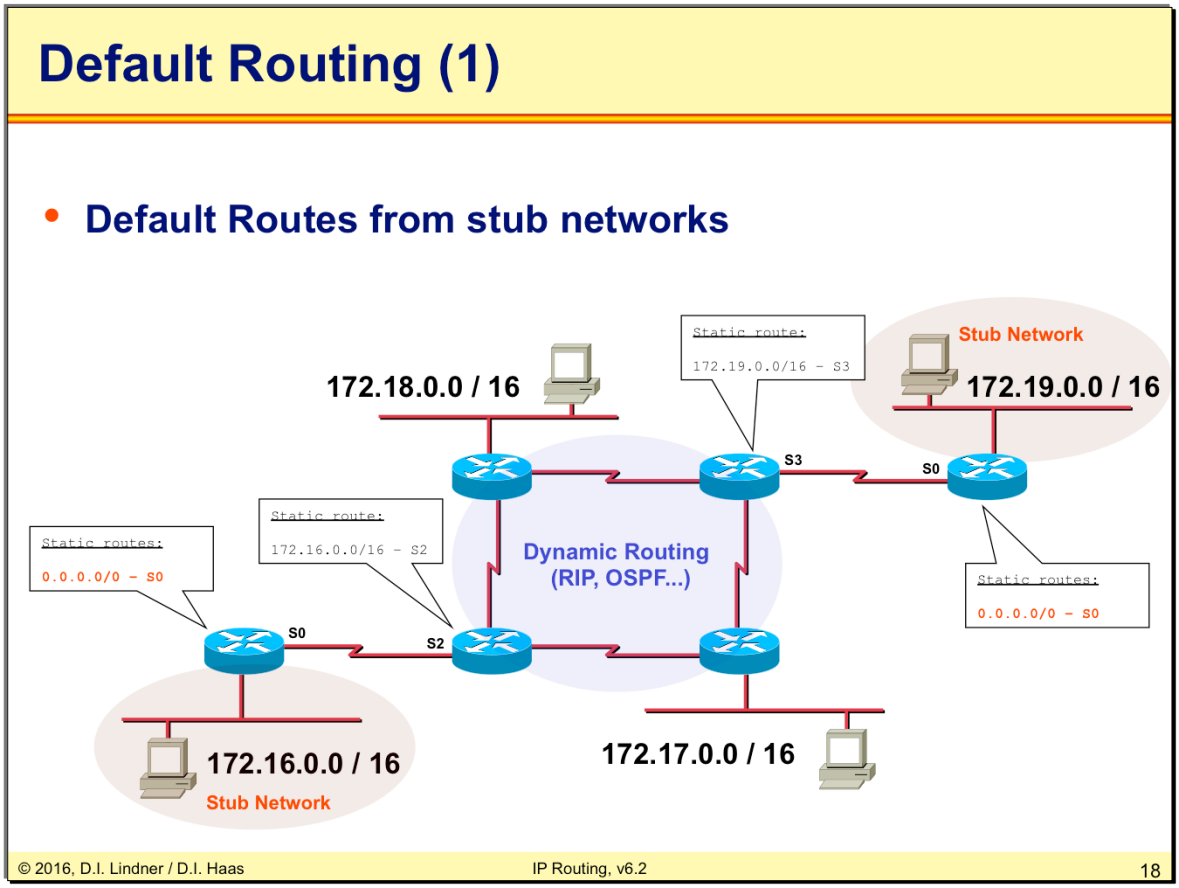
Default Route (DR)

- **Special static route in a router**
  - Traffic to unknown destinations are forwarded into the direction specified by the default route
  - Pointing to "Gateway of Last Resort“
- **In routing tables and in certain routing updates**
  - The default route is marked "0.0.0.0 0.0.0.0“
- **Hopefully, next router knows more about destination networks**
  - DR implies that another router might know more networks
- **Advantage: Smaller routing tables!**
  - DR permits routers to carry less than full routing tables

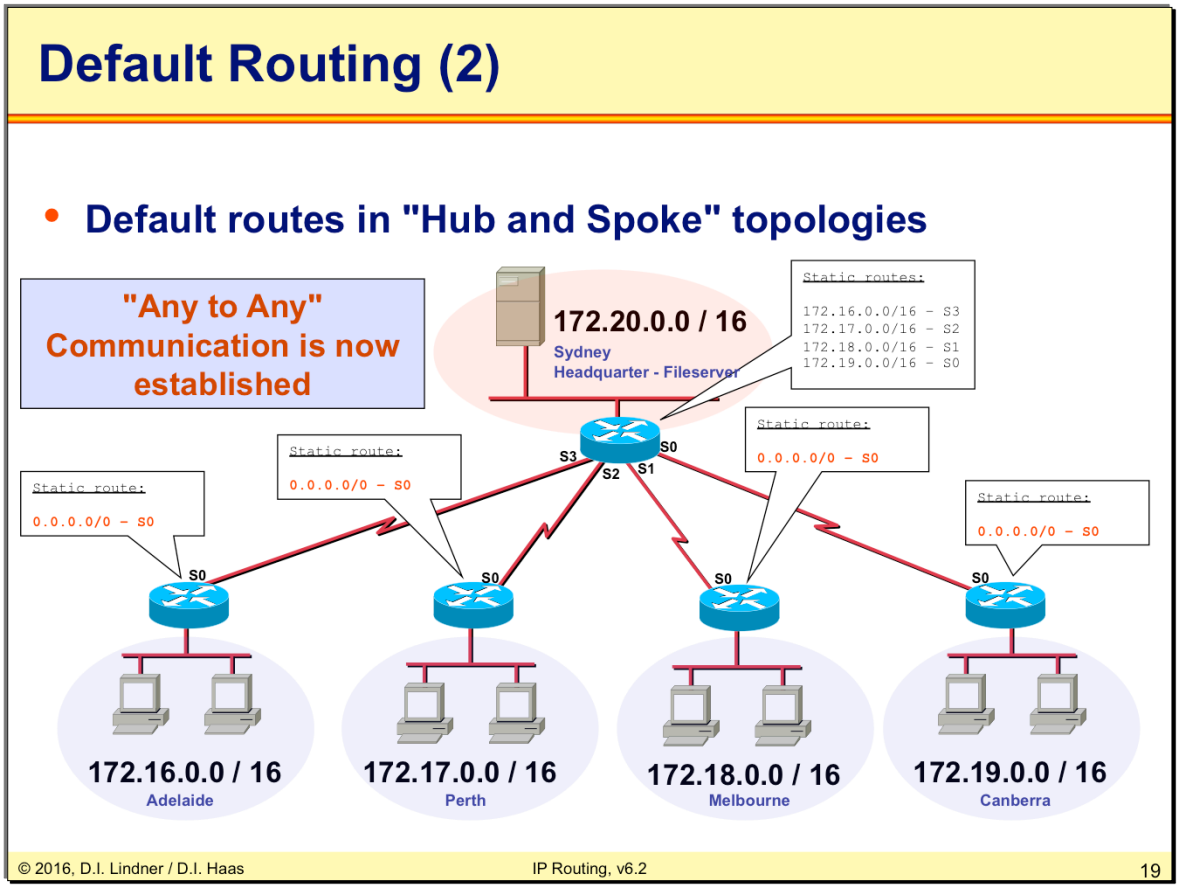
 IP Routing, v6.2 17

To get smaller routing tables there is the so called default route. When a router receives a datagram, and when the router couldn’t find the destination address of the datagram in his routing table he is forward this datagram towards his default route, hopefully the next router knows more.

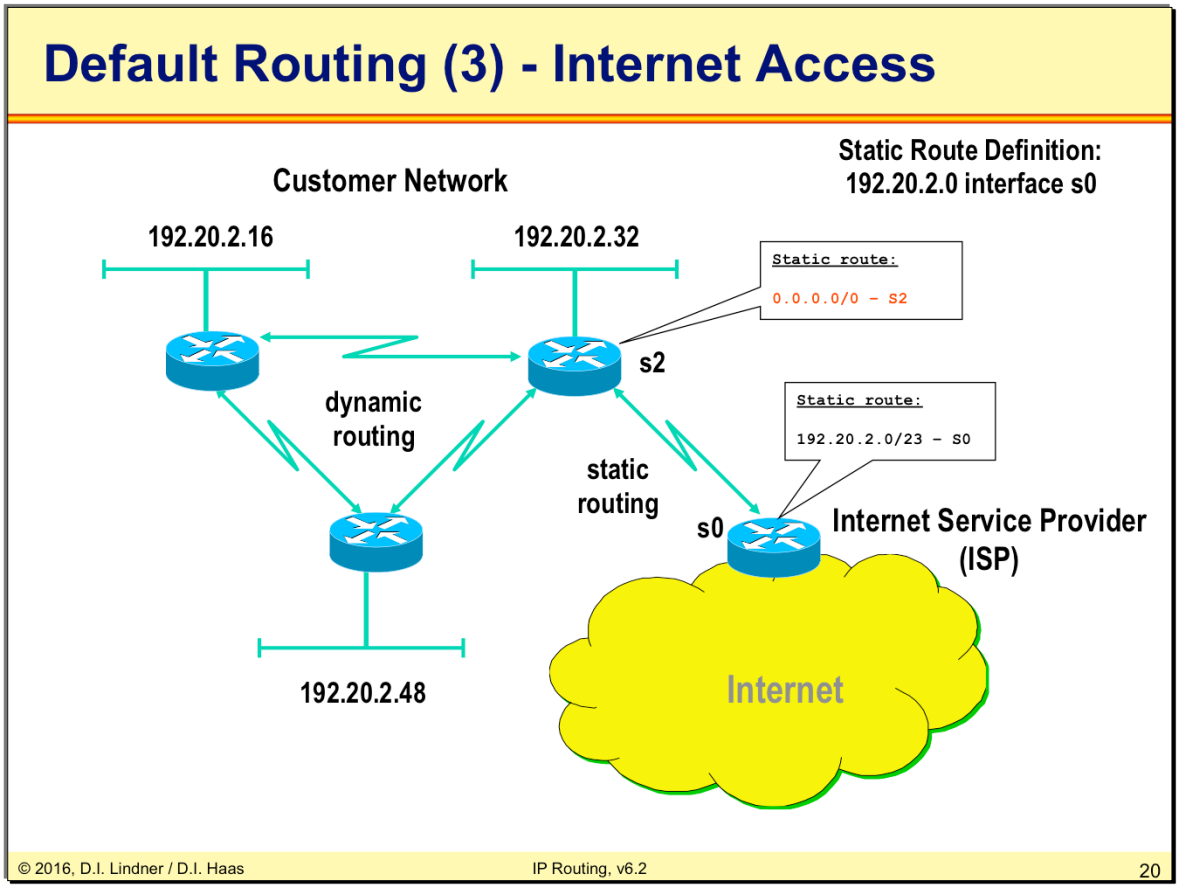
Remember the general routing principle if there is no default route known: Traffic to destinations that are unknown to the router will be discarded by the router and an ICMP message (destination unreachable) will be sent).



In this example you see the work of default routing. Every packet a router of a stub network receives will be forward to the next router, doesn't matter what destination address. The tradeoff for default route in comparison to specifying all static routes in detail is that now all unknown traffic will be forwarded to the core network in the picture even if in the core the destination network of that traffic is not known, too. So bandwidth on the WAN link to the core is wasted by such unknown traffic whereas in the other scenario unknown traffic will not leave the stub network.



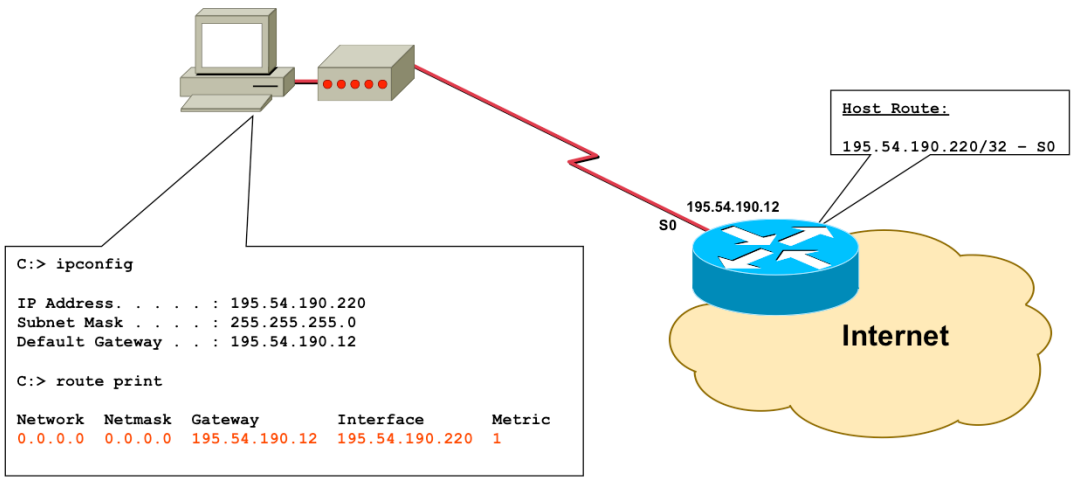
With default routing it is now possible that every branch office can talk with each other, and not only with the headquarter.



Without default route technique all routers of a customer network would need to have all routes known in the Internet their routing tables. By the time being that are 415000 routes (may 2012).

Default Routing (4)

- **Default Routes to the Internet**



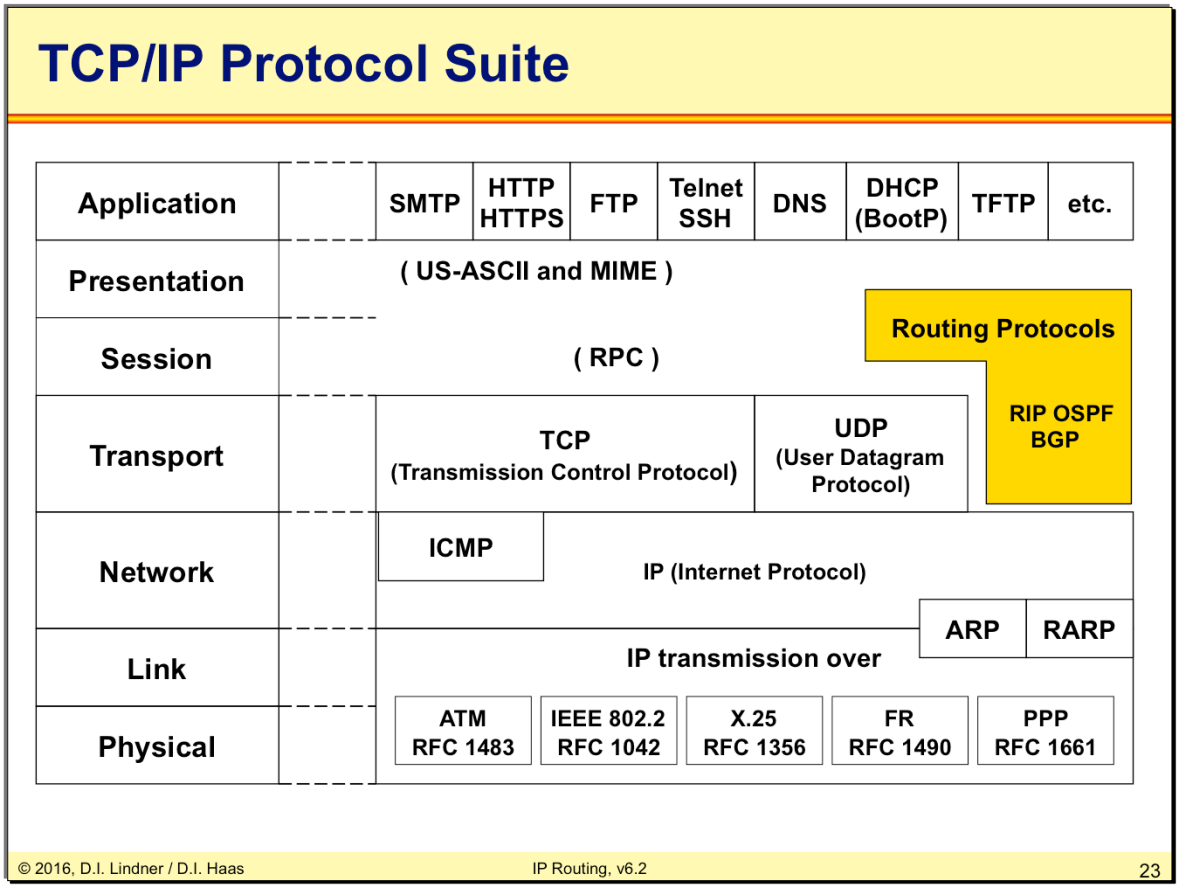
Also your home pc uses the default route.

Router IP Address 195.54.190.12

Once the host dials in, the router assigns an IP-Address (195.54.190.220) and a default gateway (195.54.190.12) to that host and also creates a "Host Route" (dynamic) that points to that host. The host takes that default gateway information and creates a default route pointing to its local interface.

Agenda

- Introduction to IP Routing
  - Basics
  - Static Routing
  - Default Route
  - Dynamic Routing
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**



RIP routing messages are carried in UDP with well known port 520.
OSPF routing messages are carried in IP with well known protocol number 89.
BGP routing messages are carried in TCP with well known port 179.

Dynamic Routing

- **Basic principle**
  - Routing tables are dynamically updated with information from other routers exchanged by routing protocols
  - Routing protocol
    - Discovers current network topology
    - Determines the best path to every reachable network
    - Stores information about best paths in the routing table
  - Metric information is necessary for best path decision
    - In most cases summarization of static preconfigured values along the given path
      - Hops, interface cost, interface bandwidth, interface delay, etc.
  - Two basic technologies
    - Distance vector, Link state

What can a dynamic routing protocol detect? Basically only loss of links and loss of routers. In case of redundancy an alternate route will be stored in the routing table.

Routing Metric

- **Routing protocols typically find out more than one route to the destination**
- **Metric help to decide which path to use**
  - Static values
    - Hop count, distance (RIP)
    - Cost like reciprocal value of bandwidth (OSPF)
    - Bandwidth (EIGRP), Delay (EIGRP), MTU
  - Variable or dynamic values
    - Load (EIGRP)
    - Reliability (EIGRP)
    - Very seldom used
      - Cisco citation:

        "If you do not know what you are doing do not even think using or touching them!"

 IP Routing, v6.2 25

Often router find more than one path to forward a packet to a given destination. The metric helps router find the "best" way. Note that there are several types of metrics used in modern routing protocols. Typically they cannot be compared with each other. For example a simple hop-count is no measure for speed (bandwidth).

Dynamic Routing

- **Each router can run one or more routing protocols**
- **Routing protocols**
  - Are information sources to create routing table
  - Announce network reachability information
    - By doing this a router declares that traffic destined to a certain network can be sent to him
    - Network reachability information flows in the opposite direction to the traffic destined to a network
- **Routing protocols differ in**
  - Convergence time, loop avoidance, maximum network size, reliability and complexity

 IP Routing, v6.2 26

In contrast to static routing where every route must be configured manually, dynamic routing works with one or more routing protocols. These protocols inform the router and create the routing table automatically. Widely used in the Internet. Convergence time is the time until all routers will have the same consistent view of the network after a topology change. Until that temporary routing loops are possible, if entries in routing tables point to each other or lead to circles.

Routing Protocol Comparison

Routing Protocol	Complexity	Max. Size	Convergence Time	Reliability	Protocol Traffic
RIP	very simple	16 Hops	High (minutes)	Not absolutely loop-safe	High
RIPv2	very simple	16 Hops	High (minutes)	Not absolutely loop-safe	High
IGRP	simple	x	High (minutes)	Medium	High
EIGRP	complex	x	Fast (seconds)	High	Medium
OSPF	very complex	Thousands of Routers	Fast (seconds)	High	Low
IS-IS	complex	Thousands of Routers	Fast (seconds)	High	Low
BGP-4	very complex	more than 100,000 networks	Middle	Very High	Low

 IP Routing, v6.2 27

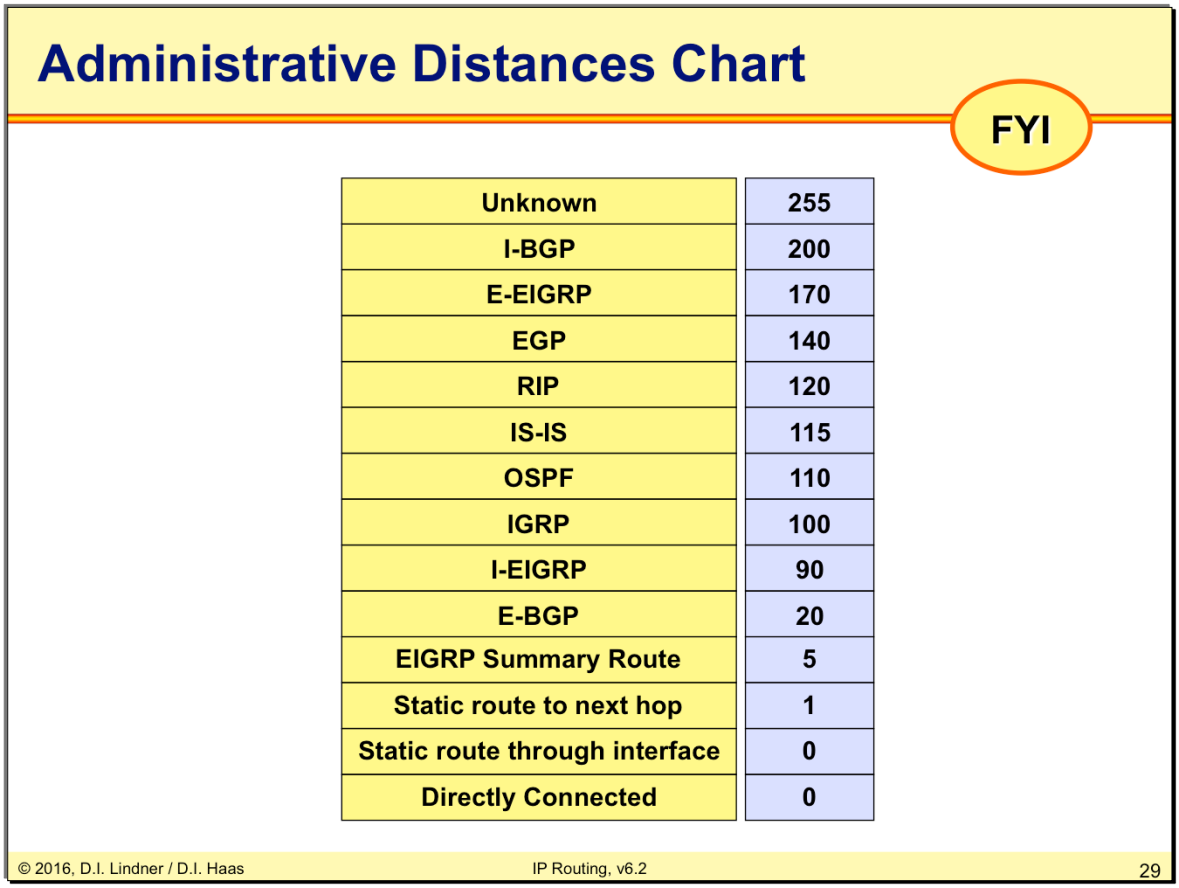
The table above gives a rough comparison of the most important routing protocols used today. Note that some values can not easily determined and are left blank for this reason.

Administrative Distance Longest Match Routing Rule

- **Several routing protocols independently find out different routes to same destination**
  - Which one to choose?
- **"Administrative Distance" is a trustiness-value associated to each routing protocol**
  - The lower the better
  - Can be changed
- **Note:**
  - If a destination network (seen in an IP datagram) matches more than one entry in the routing table
  - Then **"Longest Match Routing Rule"** is used and the best match will be taken
    - Best means the highest amount of bits from left to right in a given IP address are identical to the routing entry

 IP Routing, v6.2 28

If several different routing protocols suggest different paths to the same destination at the same time, the router makes a trustiness decision based on the "Administrative Distance", which is a Cisco feature. Each routing protocol has assigned a static AD value indicating the "trustiness" – the lower the better. Of course these values can be manipulated for special purposes.



Note the difference between static routes, if the next hop either points to an interface (AD=1) or if the route is configured as directly connected (AD=0)

AD also tells the router that E-BGP updates are more trustworthy than I-BGP messages.

Remember:

1) Using the METRIC *one* routing protocol determines the best path to a destination.

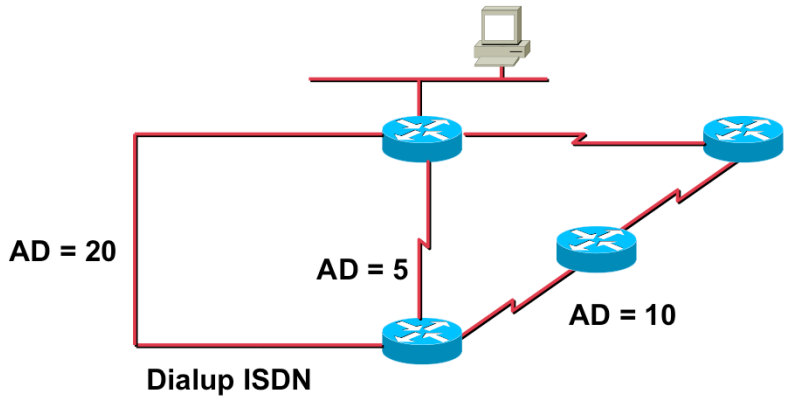
2) A router running multiple routing protocols might be told about multiple possible paths to same destination.

3) Here the METRIC cannot help for decisions because *different type of METRICS* cannot be compared with each other.

4) A router chooses the route which is proposed by the routing protocol with the *lowest* ADMINISTRATIVE DISTANCE

AD with Static Routes

- Each static route can be given a different administrative distance
- This way fall-back routes can be configured



In the example above, there are several static routes to same destination. There are three paths with different quality (more or less hops, BW, dial-up link ...). So every path has assigned a different AD. If there are problems with the main path (AD 5) the router automatically change to the next path (AD 10) and so on.

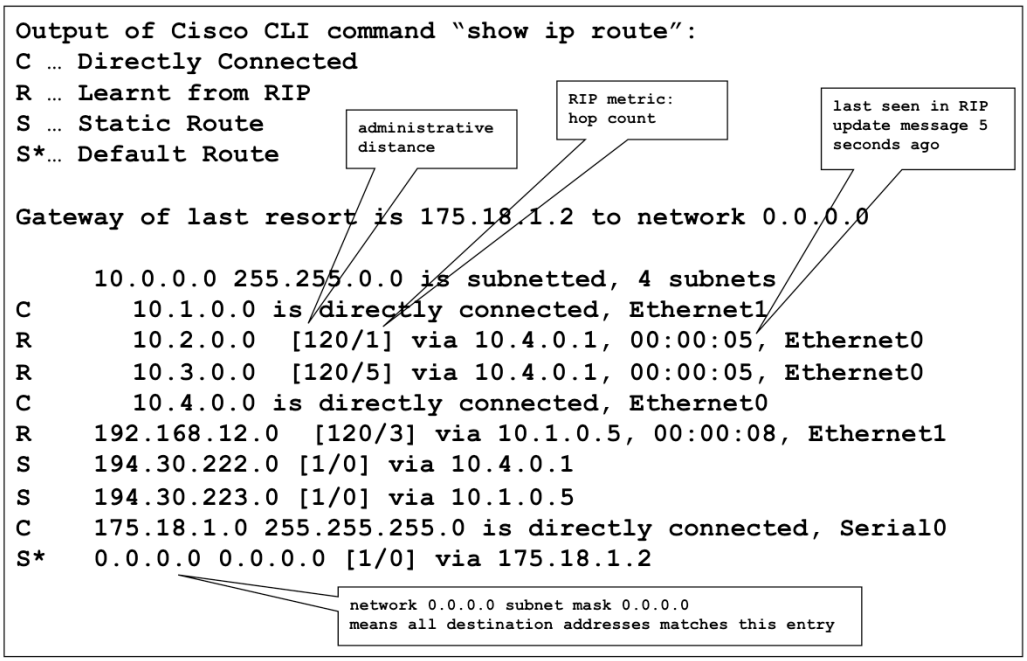
Classification of Routing Protocols

- **Depending on age:**
  - Classful (no subnet masks)
    - Routing updates carries IP net-ID only
  - Classless (VLSM/CIDR supported)
    - Routing updates carries IP net-ID and subnet mask
    - Very often prefix/length notation is used !!!
- **Depending on scope:**
  - IGP (inside an Autonomous System)
  - EGP (between Autonomous Systems)
- **Depending on algorithm:**
  - Distance Vector (Signpost principle)
  - Link State (Roadmap principle)
  - Hybrid (mixture of distance vector and link state)

 IP Routing, v6.2 31

All routing protocols can be classified three-fold. If routing protocols are able to carry a subnet mask for each route we call them "classless", otherwise "classful". Today, most modern routing protocols are classless and therefore support VLSM and CIDR. If routing protocols are used inside an autonomous system we call it "Interior Gateway Protocol (IGP)", while only "Exterior Gateway Protocols (EGPs)" are used between autonomous systems. Technically, all routing protocols use one of two possible algorithms: "Distance Vector" protocols rely on the signpost principle, while "Link State" protocols maintain a road-map for the whole network.

Routing Table Example



 IP Routing, v6.2 32

In the picture above there is example of a routing table. 0.0.0.0 is used for default gateway. The single letters at the beginning of each entry indicates how the routes were learned, for example "C" corresponds to "Directly Connected", "R" means "learned by RIP", "S" means "static route", and so on. The numbers in the brackets denote the administrative distance and the metric. For example [120/5] means AD=120, metric=5.

Distance Vector Protocols (1)

- **After powering-up each router only knows about directly attached networks**
- **Routing table is sent periodically to all neighbor-routers**
- **Received updates are examined, changes are adopted in own routing table**
  - Changes announced by next periodic routing update
- **Metric information is based on hops (distance between hops)**
  - Hop count metric is a special case for the more generic distance value between two routers
  - Hop count means distance = 1 between any two neighboring routers
- **"Bellman-Ford" algorithm**

Distance vector protocols works with the Signpost principle. A Part of the own routing table is sent periodically to all neighbor routers (e.g.: RIP: every 30 seconds).

A signpost carries the Destination network, the Hop Count (metric, "distance") and the Next Hop.

After a router receives a update, he extracts new information's. Known routes with worse metric are ignored.

Distance Vector Protocols (2)

- **Limited view of topology**
  - Next hop is always originating router
  - Topology behind next hop unknown
  - Signpost principle
- **Loops can occur!**
- **Additional mechanisms needed**
  - Maximum hop count
  - Split horizon (with poison reverse)
  - Triggered update
  - Hold down
  - Route Poisoning

Routers view is based on its routing table only. There is an exact view how to reach local neighbors but the network topology behind neighbors is hidden. Therefore such a router has only a limited view of the network topology which causes several problems. Additional mechanism are necessary first to solve problems like count to infinity and routing loops and second to reduce convergence time. That is the time to reach consistent routing tables in all routers after a topology change.

Distance Vector Protocols (3)

- **Examples**
  - RIP, RIPv2 (Routing Information Protocol)
  - IGRP (Cisco, Interior Gateway Routing Protocol)
  - IPX RIP (Novell)
  - AppleTalk RTMP (Routing Table Maintenance Protocol)

Link State Protocols (1)

- **Each two neighbored routers establish adjacency**
- **Routers learn real topology information**
  - Through "Link State Advertisements (LSAs)"
  - Stored in database (Roadmap principle)
- **Routers have a global view of network topology**
  - Exact knowledge about all routers, links and their costs (metric) of a network
- **Updates only upon topology changes**
  - Propagated by *flooding* of LSAs (very fast convergence)

 IP Routing, v6.2 36

Topology changes (link up or down, link state) are recognized by routers responsible for supervising those links and are flooded by responsible routers to the whole network again by using (**L**ink **S**tate **A**dvertisements, LSAs).

Flooding is a controlled multicast procedure to guarantee that every router gets corresponding LSA information as fast as possible but with avoiding a LSA broacast storm in case of redundancy.

Link State Protocols (2)

- **Routing table entries are calculated by applying the Shortest Path First (SPF) algorithm on the database**
  - Loop-safe
  - Only the lowest cost path is stored in routing table
  - But alternative paths are immediately known
  - Could be CPU and memory greedy
    - Mainly a concern in the past
- **Large networks can be split into areas**

Applying the SPF algorithm on the link state database, each router can create routing table entries by its own.

Link State Protocols (3)

- **With the lack of topology changes**
  - Local hello messages are used to supervise local links (to test reachability of immediate-neighboring routers)
  - Therefore less routing overhead concerning link bandwidth than periodic updates of distance vector protocols
- **But more network load is caused by such a routing protocol**
  - During connection of former separated parts of a network
  - During topology database synchronization

Link State Protocols (4)

- **Examples**
  - OSPF (Open Shortest Path First)
  - Integrated IS-IS (IP world)
    - note: Integrated IS-IS takes another approach to handle large networks (topic outside the scope of this course)
  - IS-IS (OSI world)
  - PNNI (in the ATM world)
  - APPN (IBM world),
  - NLSP (Novell world)

Summary

- **Routing is the "art" of finding the best way to a given destination**
- **Can be static or dynamic**
  - Static means: YOU are defining the way packets are going
  - Dynamic means: A routing protocol is "trying" to find the best way to a given destination
- **In today's routers the route with the longest match is used**
- **Routing protocols either implement the principle *Distance Vector* or *Link State***

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

RIPv1 - Routing Information Protocol

- **Interior Gateway Protocol (IGP)**
  - Due to inherent administrative overhead traffic, RIP suits best only for smaller networks
  - Routing decisions are based upon hop count measure
- **Distance-Vector Routing Protocol**
  - Bellman Ford Algorithm
  - RFC 1058 released in 1988
- **Classful**
  - No subnet masks carried
- **RIPv1 was initially released as part of BSD 4.2 UNIX**
  - Hence RIP got wide-spread availability
- **RIPv1 is specified in RFC 1058**
  - RFC category „historic“

 IP Routing, v6.2 42

RIP is a so-called distance vector routing protocol – its routing updates are like "signposts" pointing to the shortest-hop path to known destination networks. The algorithm has been developed by R. E. Bellman, L. R. Ford, and D. R. Fulkerson and has first been implemented in the ARPANET in 1969. In the mid-1970s, Xerox created the "Gateway Information Protocol" (GWINFO) to route the Palo Alto Research Center (PARC) Universal Protocol, also known as "PUP". PUP became the Xerox Network Systems (XNS) protocol suite and GWINFO became XNS RIP. And XNS-RIP was the basis for Novell's IPX RIP, AppleTalk's Routing Table Maintenance Protocol (RTMP), and IP RIP. We will only discuss IP RIP here.

RIP is an Interior Gateway Protocol (IGP), that is, RIP is only used inside an Autonomous System (AS). Further explanations about AS and IGP are given in the BGP part of this chapter.

RIP is an classful routing protocol, because RIP (version 1) does not bind subnet-masks to the routes. So RIP (version 1) assumes classful addressing. Subnet masks can be used as long as discontigous subnetting is avoided.

RIP Basics

- **Signpost principle**
  - Own routing table is sent periodically (every 30 seconds)
- **What is a signpost made of ?**
  - Destination network
  - Hop Count (metric, "distance")
  - Next Hop ("vector", given implicitly by sender's address!)
- **Receiver of update extracts new information**
  - New is information about a network either not known so far or an already known network with a better metric
  - Already known routes with worse metric are ignored
  - Adapts the routing table and again sent periodically its routing table

 IP Routing, v6.2 43

The whole distance vector philosophy is based upon the signpost principle – each router sends periodically a copy of his own routing table to each neighbor. Upon receiving such routing update, a router extracts unknown routes or routes that improved in metrics. For RIP the update period is 30 seconds.

Using this principle, each router learns how to reach destinations only via signpost – the routing details along the path are unknown. The routing update (signpost) basically consists of a list of destination networks and hop counts ("distances") associated to it. For all these destinations there is only one next hop: the sending router's address.

"Routing By Rumor"

- **Good news propagate quickly**
  - 30 seconds per network
- **Bad news are ignored**
  - Except when sent by routers from which these routes had been learned initially
  - But better news from ANY router will be preferred
- **A network disappears from the routing table**
  - If not refreshed within180 seconds by some routing updates
- **Hence unreachability of networks is propagated very slowly**
  - At least 180 seconds

 IP Routing, v6.2 44

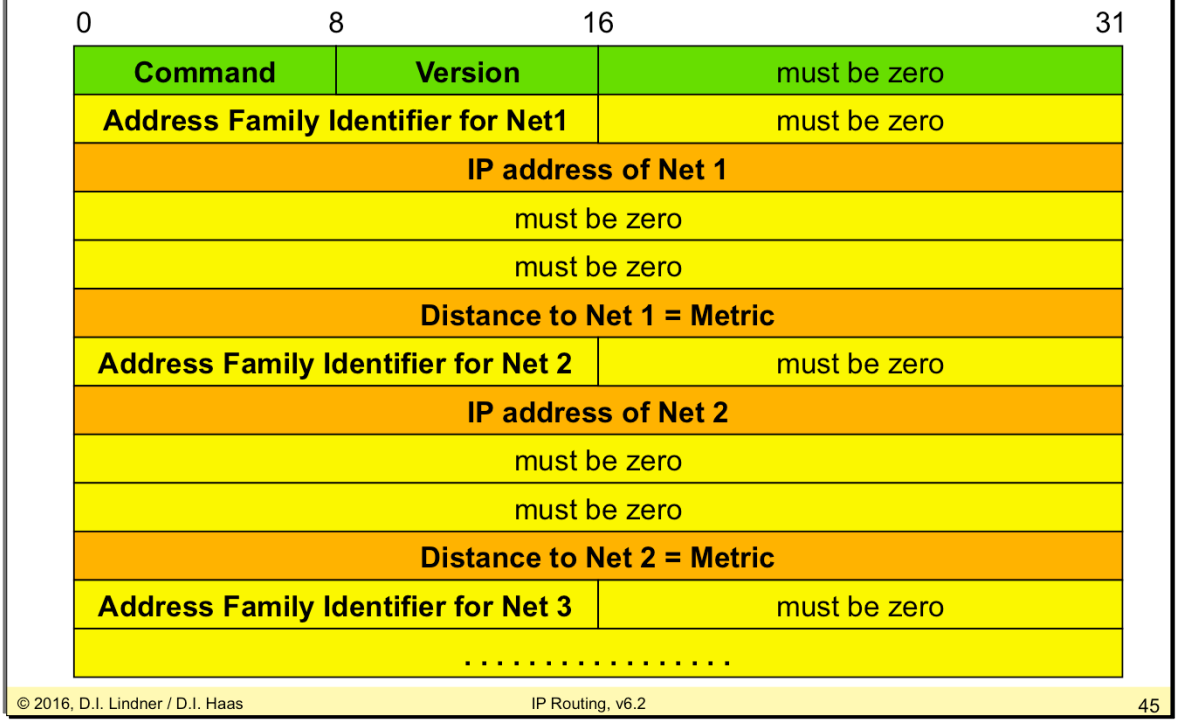
Bad news (= network reachabilities with worse metric) are only accepted if this message has been sent by that router from which we previously learned about that route.

Since RIP should discover the best routes to each destination, any routing update is accepted that contains a better route than previously learned.

A route is declared unreachable if not being refreshed by routing updates during 180 seconds.

In the worst case "bad news" propagate very slowly through the network. Special unreachable-messages have been introduced later in order to improve the convergence time. Unreachable messages are normal routing updates but with metric set to "infinity".

RIPv1 Message Format



The RIP version 1 Message format simply consists of a header, indicating command-type and version, and a number of sections reflecting a routing table entry. Up to 25 route entries per packet are allowed. Note that each route does not include a "next-hop" address! The next-hop address is simply the source address of this packet, that is, the originator declares himself as next-hop for all listed routes. Also note that there are several fields reserved as "Must be zero" to leave space for future improvements. We will see, that RIPv2 uses these fields.

RIP message fields:

Command:

- Request (1): router or IP host requests for a routing update
- Response (2): response to a request but also used for periodic routing-updates

Version: version number of the RIP protocol (= 1 for RIP)

Address Family Identifier: Because RIP was not build for IP only (AFI in the case of IP -> 2)

IP address of Net x: IP-address of the announced network x

Distance to Net x = Metric = Hop-count to net x

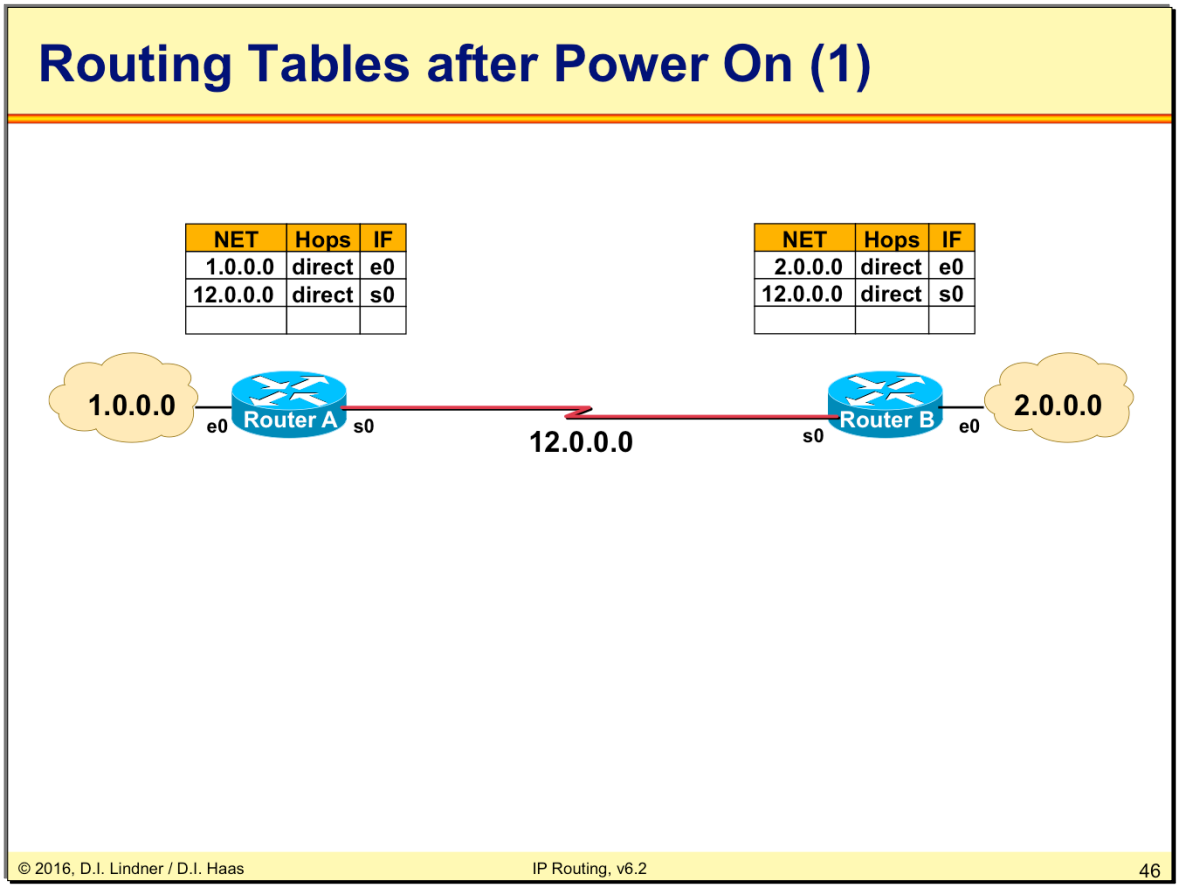
Note that a request is for specific entries (i. e. not for the whole table), the requested information is returned in any case, that is no split horizon is performed and even subnets are returned if requested. If there is exactly one entry in the request, with an address family identifier of zero and a metric of infinity (16), this is a request to send the entire routing table.

RIP message is sent within UDP payload

- UDP Port 520, both source and destination port
- Maximum message size is 512 bytes

L2 Broadcast + IP Limited Broadcast

- Because we do not know neighbor router addresses
- On shared media one update is sufficient



RIP in detail (1):

After booting the non-volatile configuration-memory tells a RIP router to which networks it is directly connected.

This information is loaded into the routing table.

Basically the routing table contains:

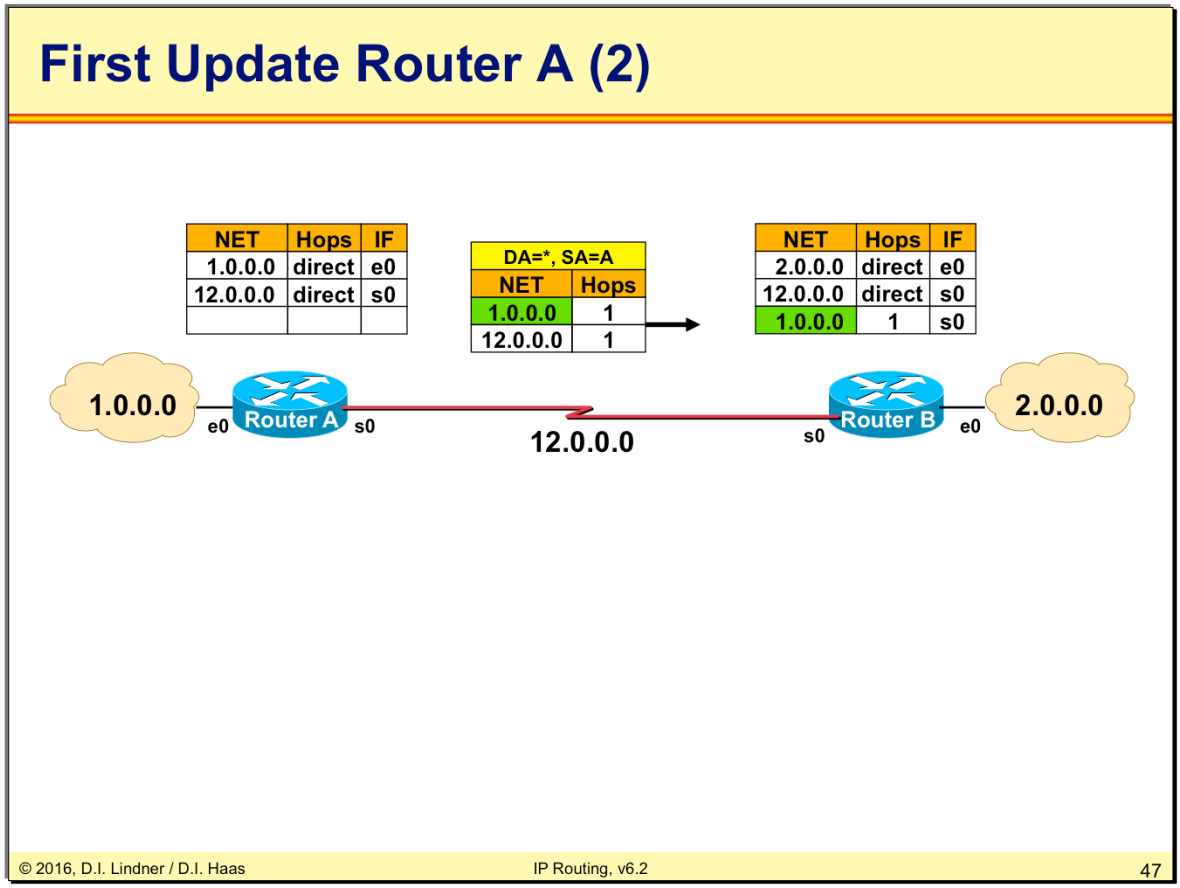
- The net-ID of the directly connected networks and the associated distance (in hops) to them.
- Directly connected networks have hop-count = 0.

This routing table is distributed periodically (every 30 seconds) to all directly connected networks = routing update.

RIP routing updates are sent as:

- Broadcast MAC-frame in case of LAN
- IP-limited-broadcast datagram
- UDP-segment with well known port number 520

Directly reachable routers receive this message, update their own routing tables, and hence generate their own routing updates reflecting any corresponding modifications.



This is the basic principal of RIP (Without Split Horizon). Every 30 seconds a router sends his whole routing table to every neighbor router and increases the Hop-Count by 1.

The router who receives this data add the new information (green in the above picture) in his routing table. If a router already knows about a better path – for example a direct connection to a net -- he will ignore this information.

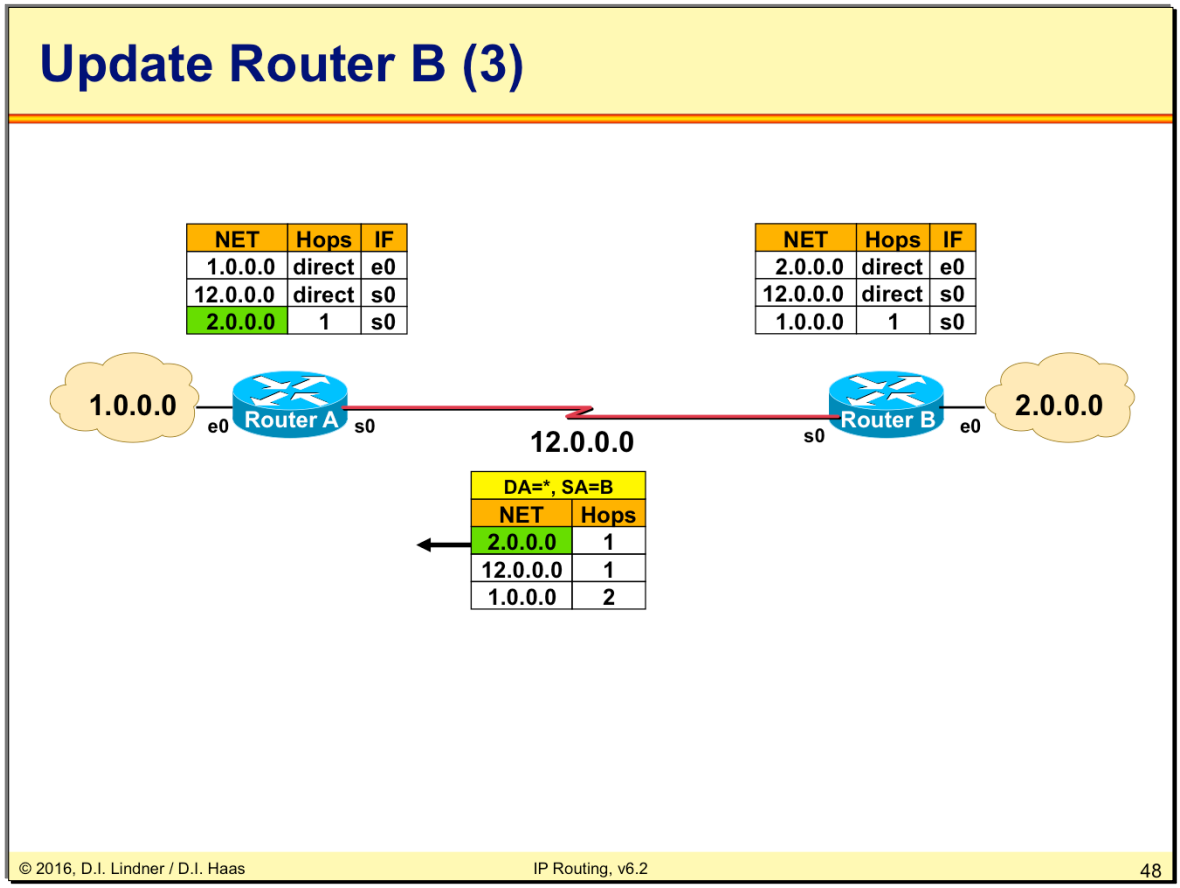
RIP in detail (2):

If a routing update tells a better metric than that one currently stored in the table:

- The routing table must be updated with this new information.
- This update does not take care about if the sender of this routing-update is also the router which is currently selected as next hop.
- “Good news“ are quickly adapted.
- Note: RIPv1 trusts good news from any source ("trusted news").

If a routing update tells a worse metric than that one currently stored in the table:

- The routing table must be updated with this new information if the sender of this routing-update is the next-hop router for this network. That is: the actual VECTOR in the table is identical with the source address of the routing-update.
- Routing-updates from other routers than that one currently registered in the table are ignored.
- Summary: routing-update with worse metric is only relevant if the comes from that router mentioned in the actual table entry.

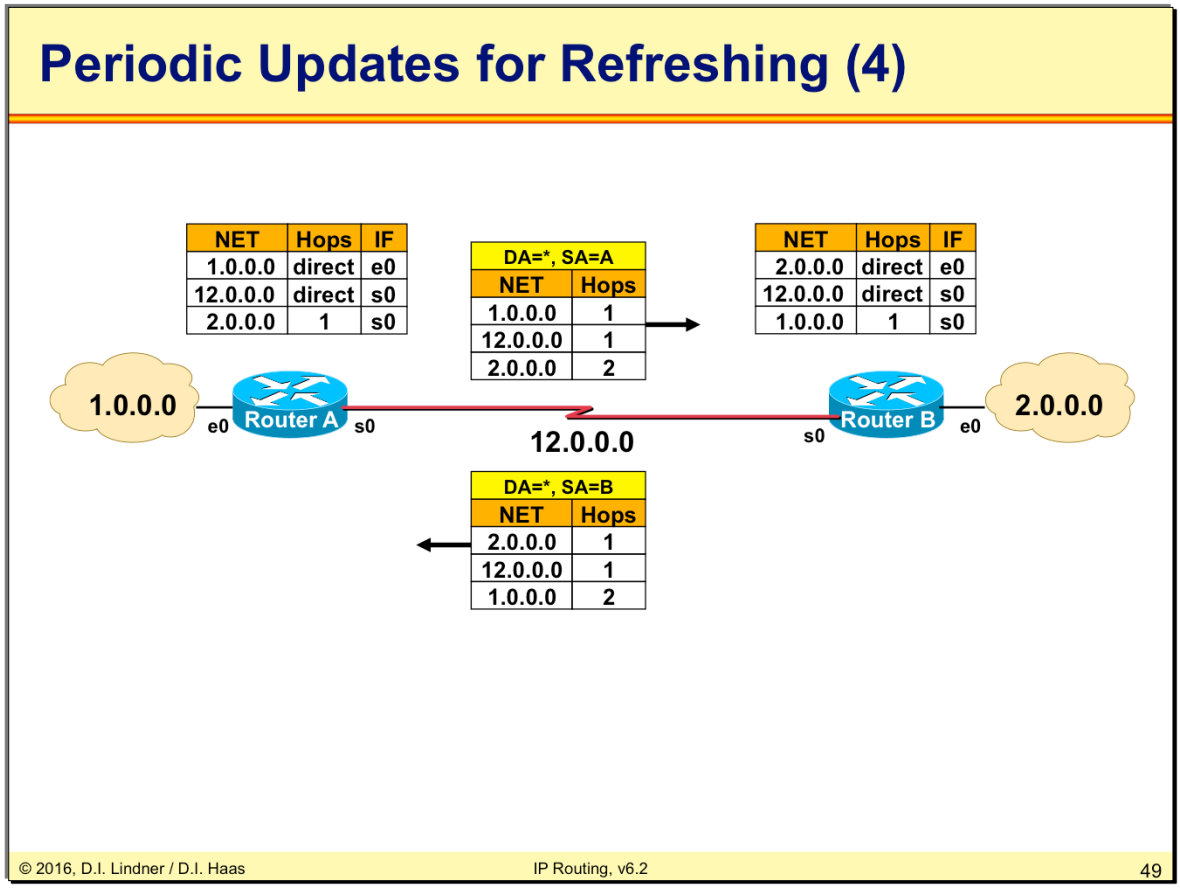


RIP in detail (3):

After some time all routers know about all network addresses of the whole network.

If different routing updates (from different routers) contain the same net-ID then there are redundant paths to this network:

- Only the path with the lowest hop-count is stored in the routing-table.
- On receiving equal hop counts, the net-ID of the earlier one will be selected.
- Hence, between each two networks exists exactly one active path.



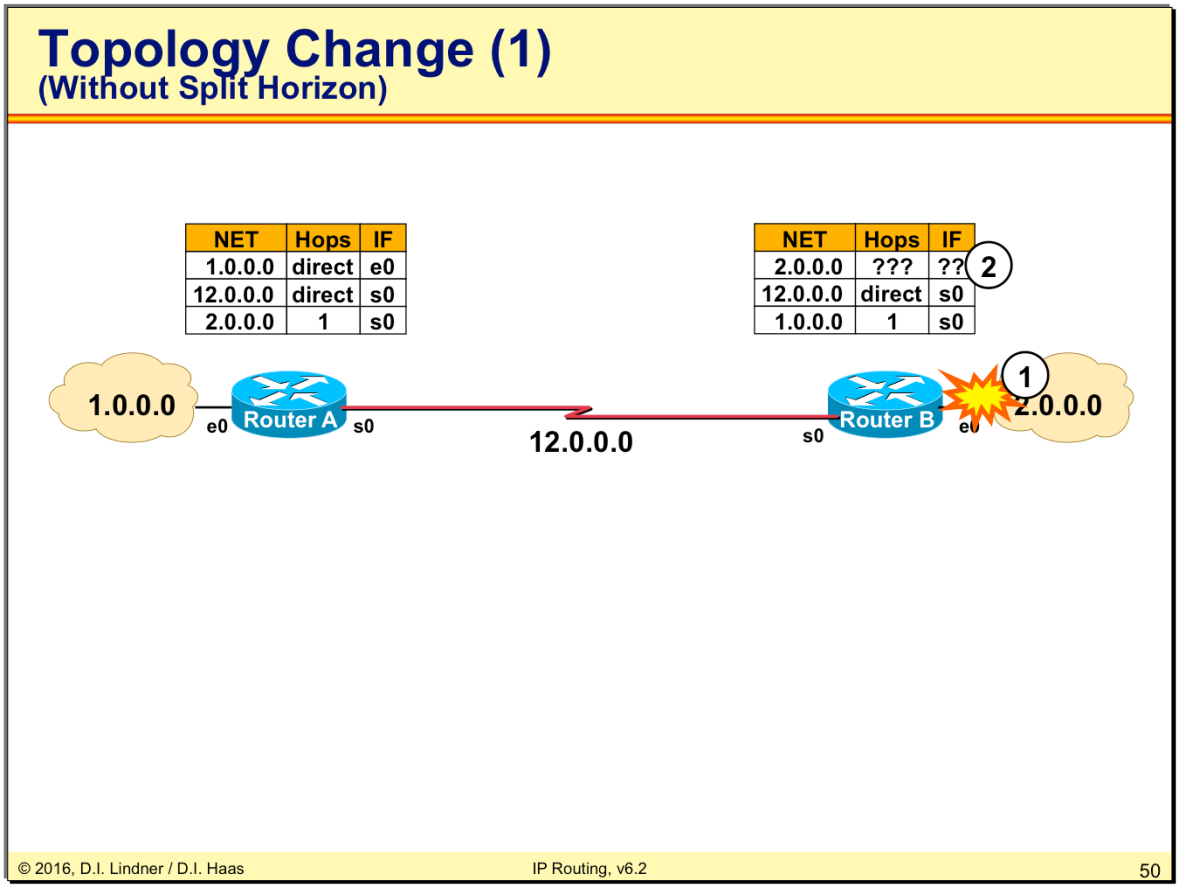
Now we have reached the stable (converged) state: all networks are known to all routers; the periodic updates are only used to refresh the routing table entries, but no new information is carried by them.

RIP in detail (4):

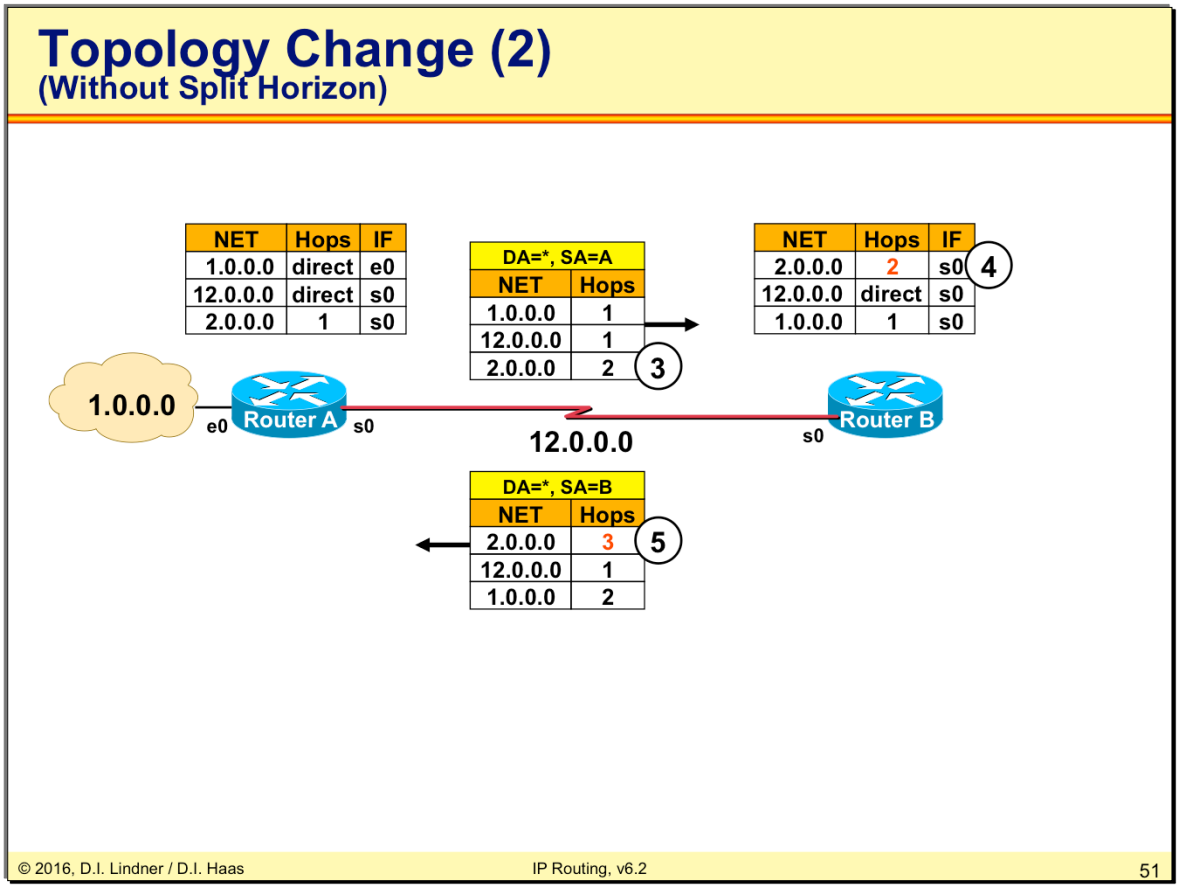
Routing tables are periodically refreshed by routing-update messages.

When a routing table entry is not refreshed within 180sec:

- This entry is considered to be obsolete.
- Possible reasons: router-failure, network not reachable.
- Without any special mechanisms we have to wait for 180sec at least before all routers have consistent routing tables again.
- Slow adaptation of "bad news“ .
- Attention: during these 180sec, forwarding of IP datagrams is still done according to the routing table.
- Improvement of convergence time is only possible by using a special network-unreachable message which is distributed to all other routers.

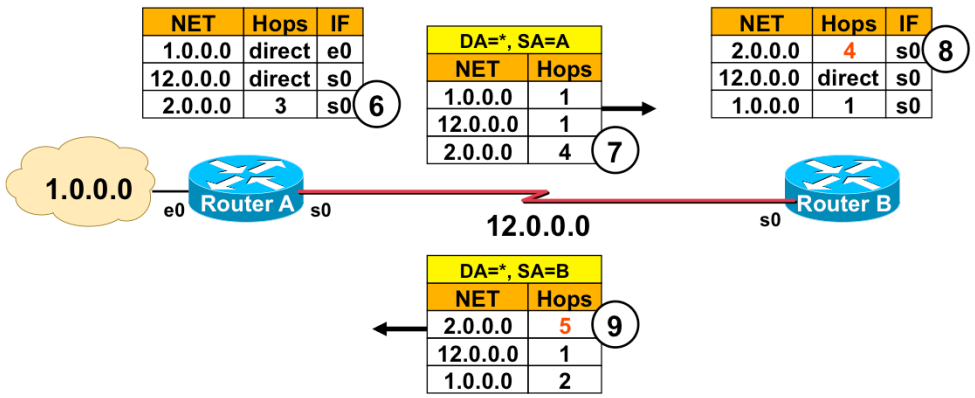


In this example we see what would happen if network 2 crashes (1). Immediately, router B has no more information about this net (2). What would happen if router A sends a routing update now?



Now router B receives a routing update from router A including reachability information about network 2 (3). Because router B has no information about network 2 he adds this information in his routing table (4) and continuous sending his normal routing updates to router A, hereby increasing the hop count by 1 (5).

Topology Change (3)
(Without Split Horizon)



...Count to Infinity...
During count to infinity datagrams to network 2.0.0.0 are caught in a routing loop

 IP Routing, v6.2 52

Router A has to adapt to the worse metric (6) and the next update will show the adapted routing table entry increased by one to router B (7). The same will happen at router B (8, 9).. Count to infinity accurse. Now update packets are caught in a routing loop. IP datagrams destined to network 2 will circle between router A and B until their TTL is decremented to 0 and they are killed by the routers to avoid endless circling. You see routing loops are not funny for the performance of a link. Image that the link may be used for IP datagrams destined to other networks not shown in drawing. These datagrams will suffer from the circling IP datagrams destined to network 2 during the routing loop situation.

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

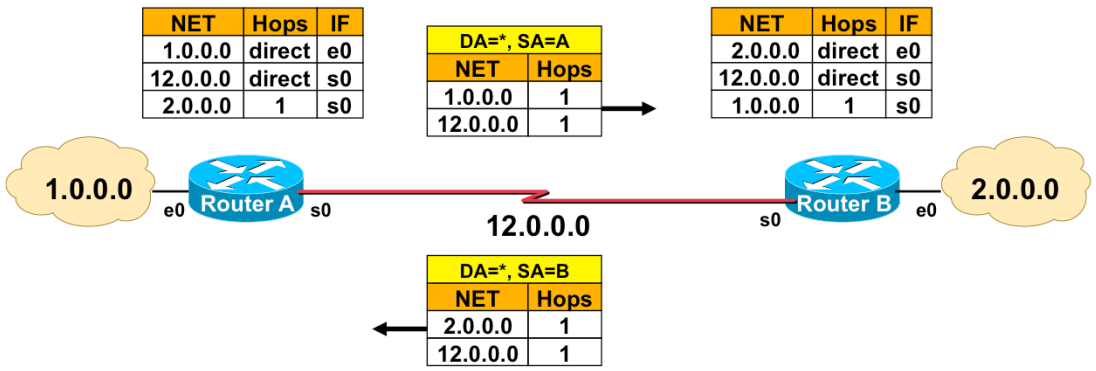
Split Horizon

- **A router will not send information about routes through an interface over which the router has learned about those routes**
  - Exactly THIS is split horizon
- **Idea: "Don't tell neighbor of routes that you learned from this neighbor"**
  - That's what humans (almost) always do:

    Don't tell me what I've told you !
- **Split horizon**
  - Cannot 100% avoid all routing loops!
- **See RIP at work with split horizon on the following slides**

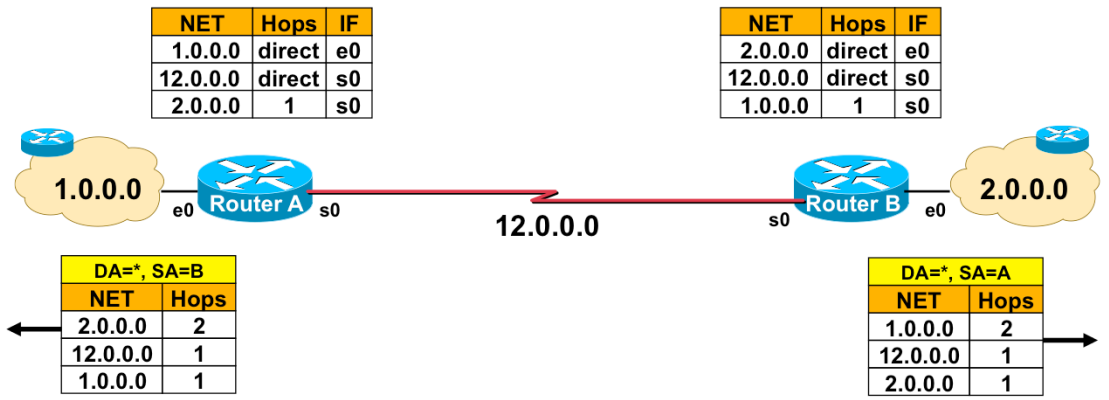
Nowadays all routers work with Split Horizon, there is now RIP-Network without it. The principle of Spilt Horizon is simple: “Don’t tell neighbor of routes that you learned from him”.

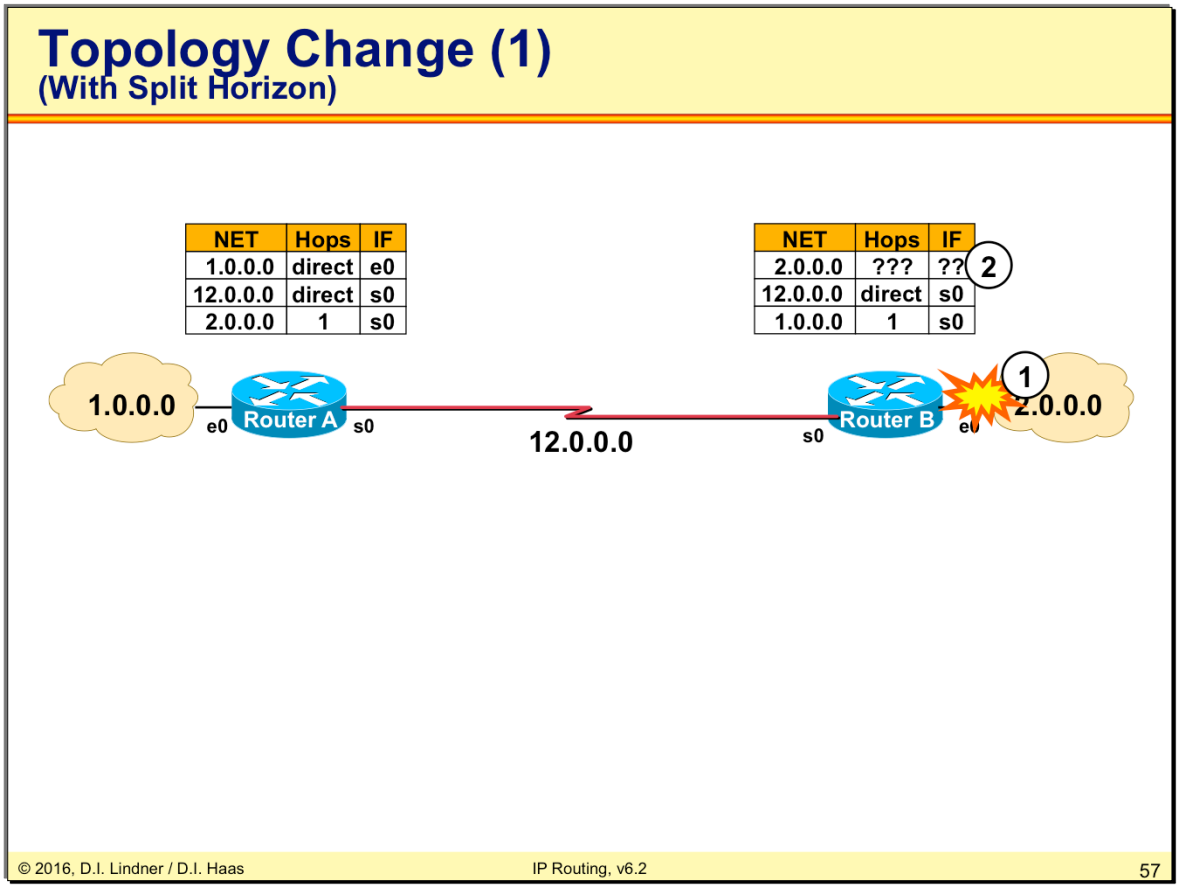
Periodic Updates With Split Horizon (1)



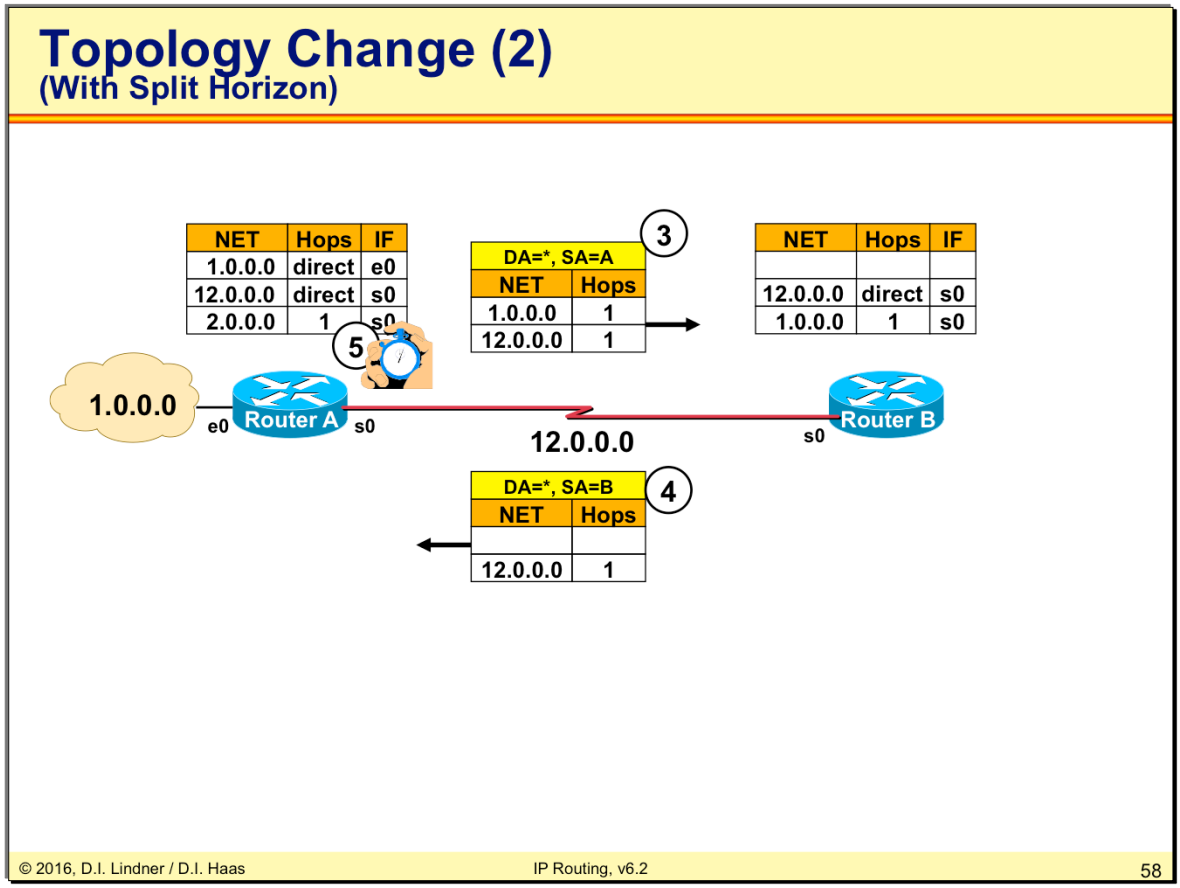
 IP Routing, v6.2 55

Periodic Updates With Split Horizon (2)

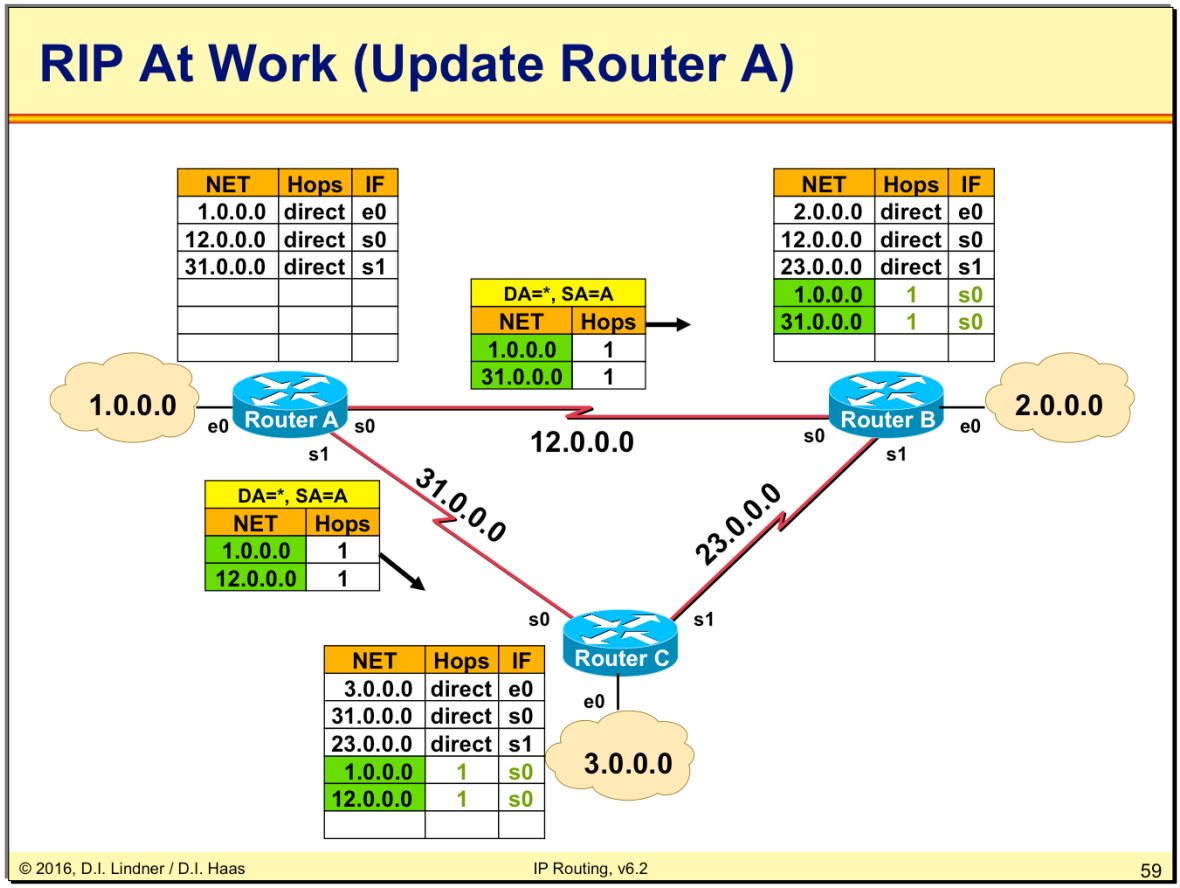




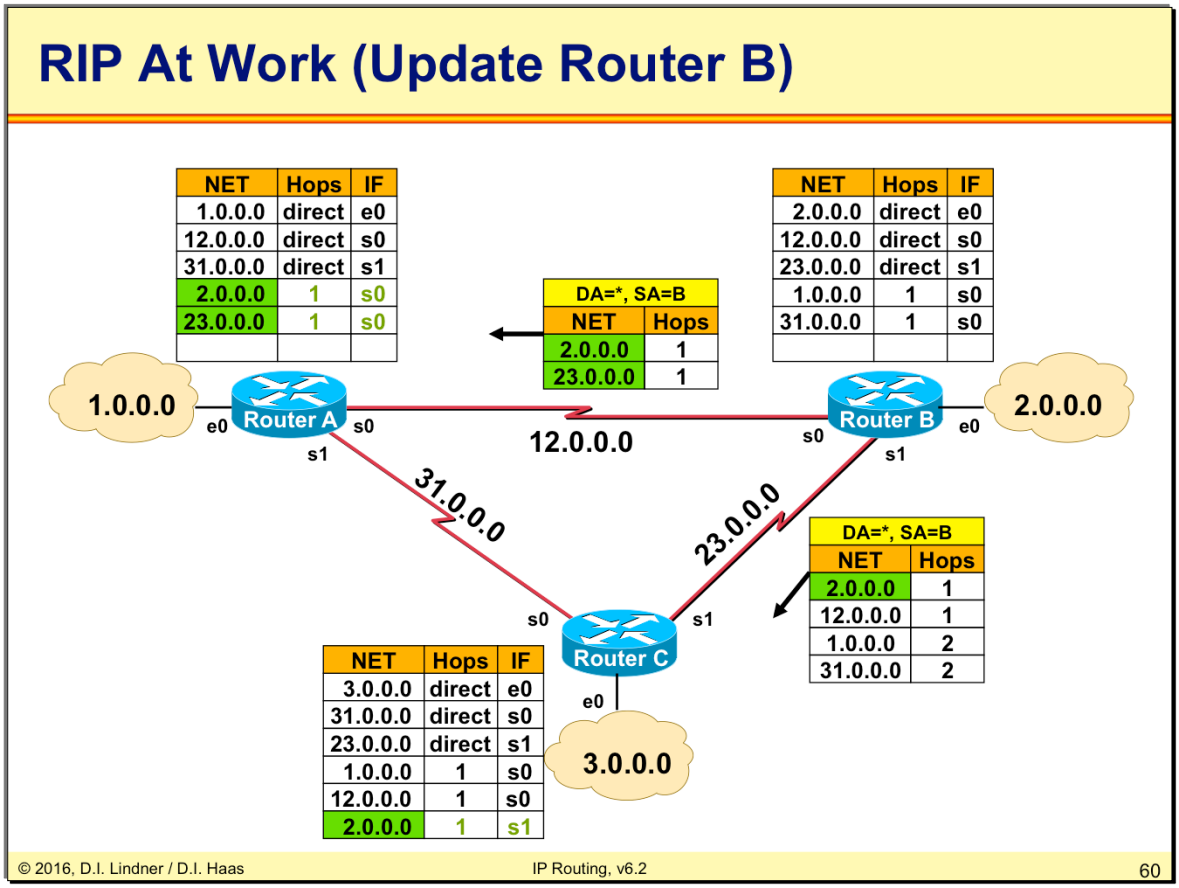
Let us see what happens when network 2 crashes (1), the routing entry in B disappears (2) and split horizon is in place.



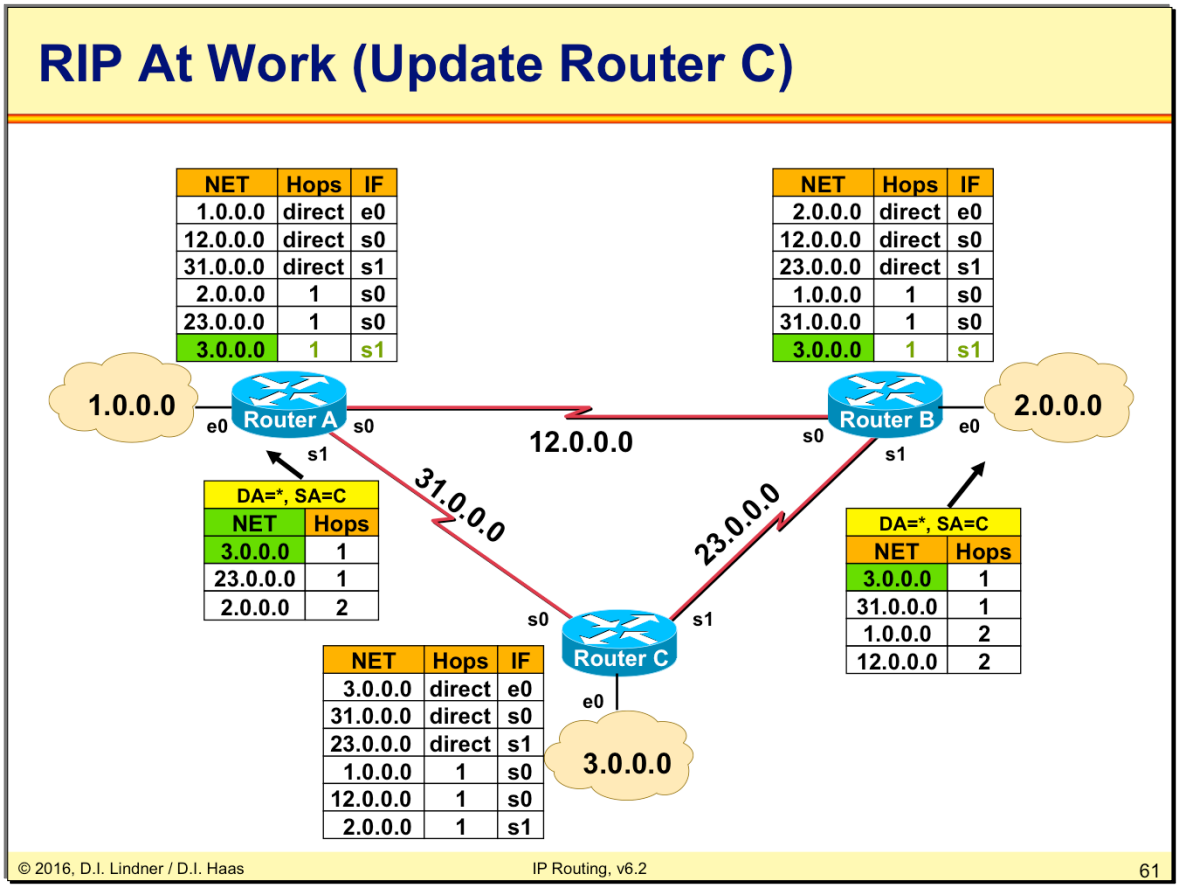
Router B receives a routing update from router A including no information about network 2 hence adapting the routing table in router B for network 2 is avoided (3) . On the other hand router B will not announce network 2 any longer (4). So the routing entry in router a will disappear after 180 seconds (5).



Split Horizon at work: Router A didn't tell router B about the network 12 and router A didn't tell router C about the network 31, because the router knows that router B must have a direct connection to network 12 and that router C must have a direct connection to network 31.



And so router B tells router A only about network 2 and 23 and router C only about network 2, 12, 1 and 31. 1 and 31 are announced by router B because they have been learnt from s0 but not from s1 hence split horizon does not apply. Of course 1 and 31 are ignored by C because of the worse metric 2. C has already learnt 1 and 31 from s0 with metric 1!



Router C do the same.

1 and 12 are announced by router C to B because they have been learnt from s0 but not from s1 hence split horizon does not apply. Of course 1 and 12 are ignored by B because of the worse metric 2. B has already learnt 1 from s0 with metric 1 and is directly connected to 12!

2 is announced by router C to A because it has been learnt from s1 but not from s0 hence split horizon does not apply. Of course 2 is ignored by A because of the worse metric 2. A has already learnt 2 from s0 with metric 1!

At the end every router knows the route to every network.

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Count To Infinity

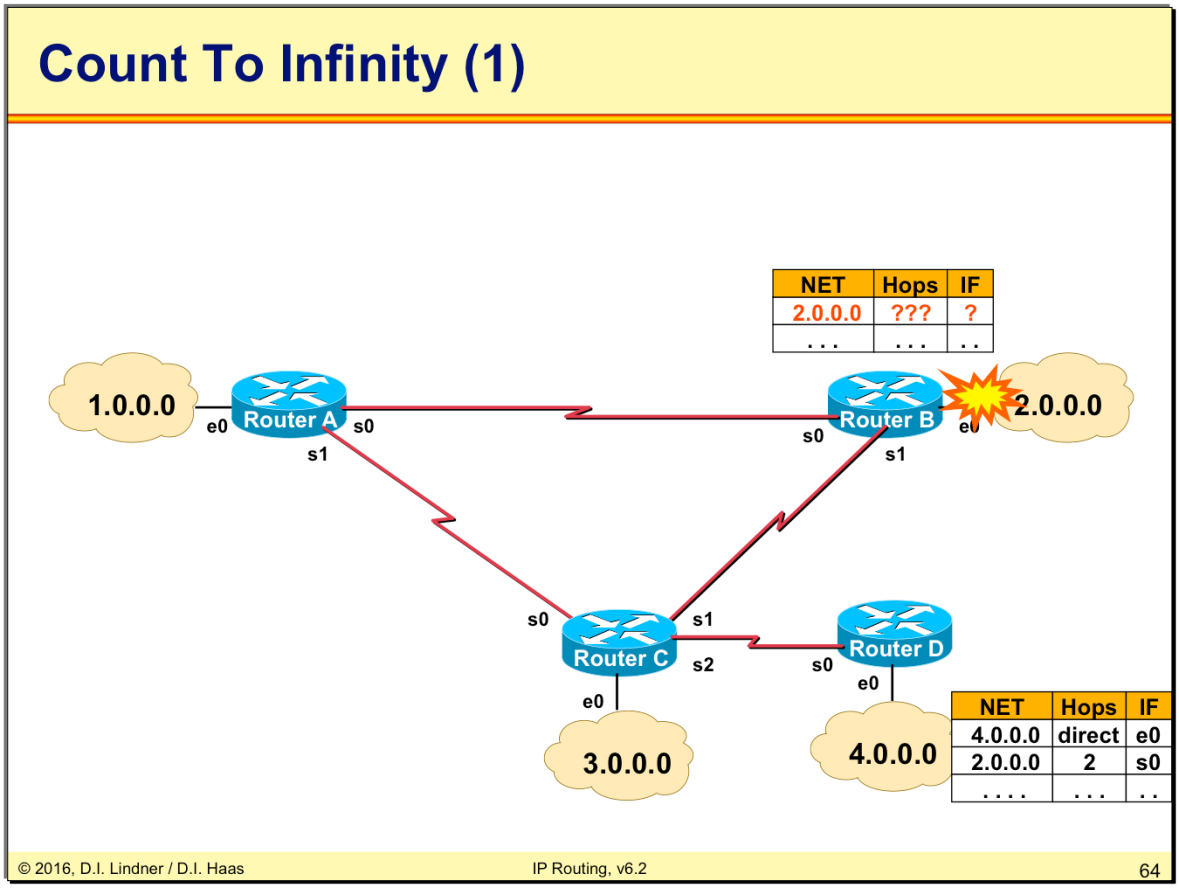
- **Main problem with distance vector protocols**
- **Unforeseeable situations can still lead to count to infinity**
  - Access lists
  - Disconnection and connections
  - Router malfunctions
  - ....
- **During that time, routing loops occur!**
- **We need an additional element**
  - Maximum Hop Count = 16 for RIP
  - Hop count =16 can also be used as unreachabilty message

 IP Routing, v6.2 63

Because of the simple principle of RIP (Distance Vector protocol), we cannot prevent Count to Infinity. Access Lists, Disconnection and connections, Router malfunction, etc. can always lead to it, there is no 100% solution.

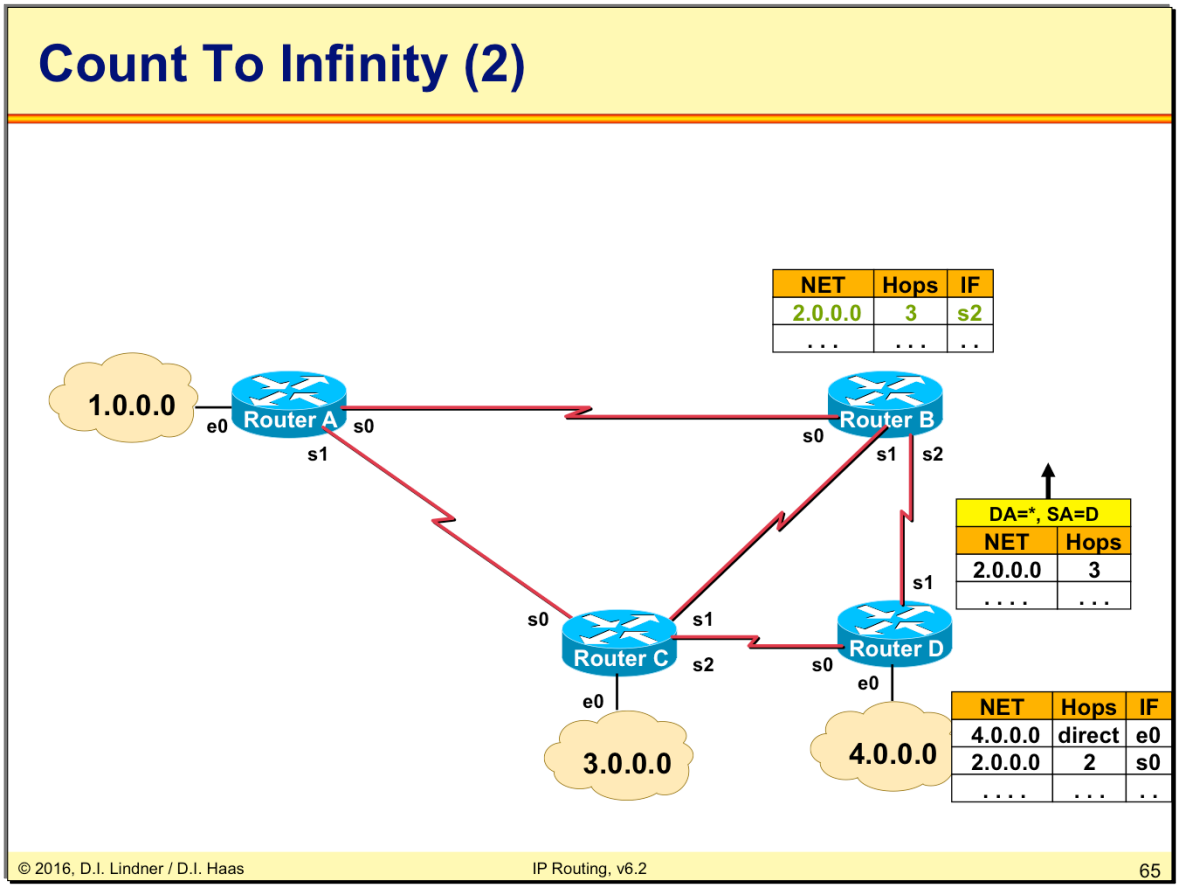
We need a more general approach to avoid that → Maximum Hop Count, that's the only failsafe solution.

Maximal distance between each two subnets is limited to 16 therefore the hop count between two end-systems cannot exceed 15. A DISTANCE-value of 16 in the routing-table means that the corresponding network is not reachable. Using hop count = 16 in a routing update allows a router to immediately indicate the failure of a network as the routers have not to age out this entry in all routing tables hence waiting at least for 180s.

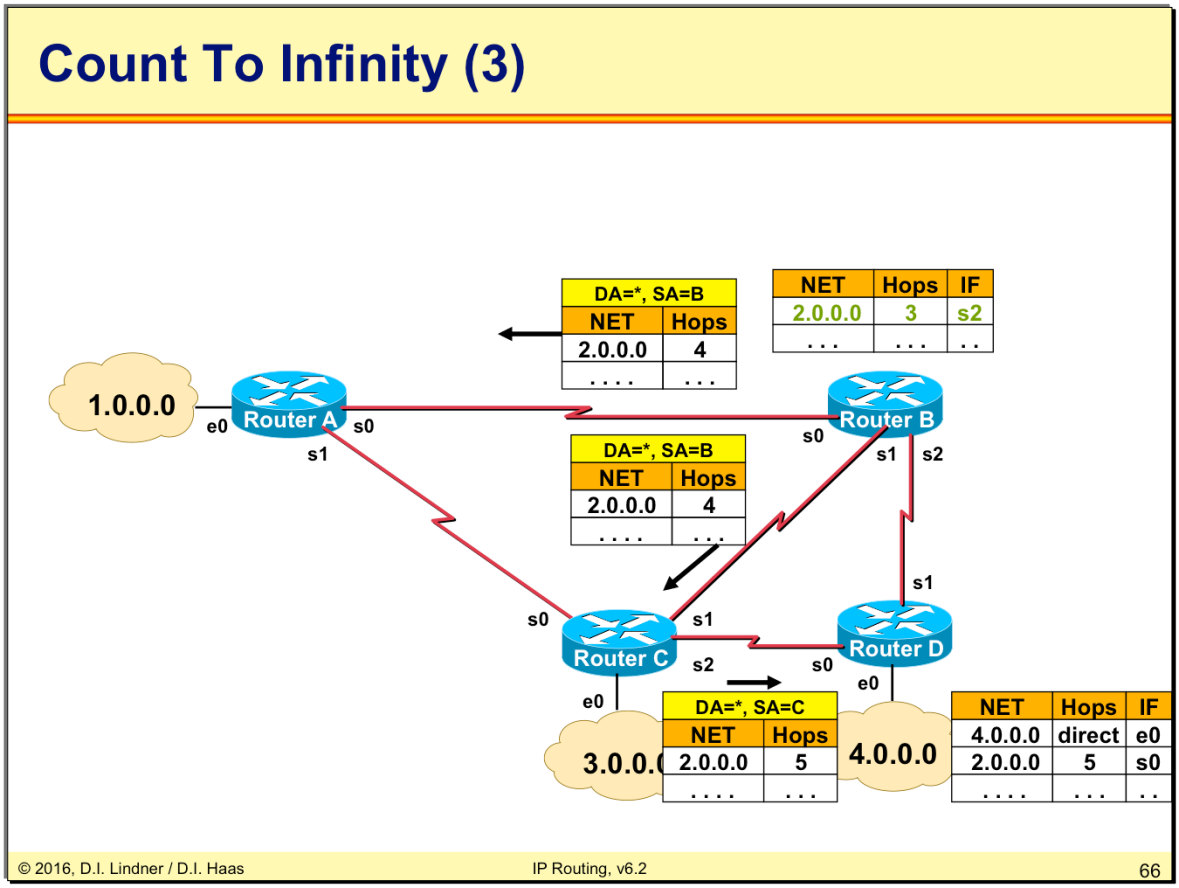


Lets us look to another example where Count to Infinity is approaching. Although Split Horizon is implemented!

We have a network with 4 routers, suddenly net 2 crash.



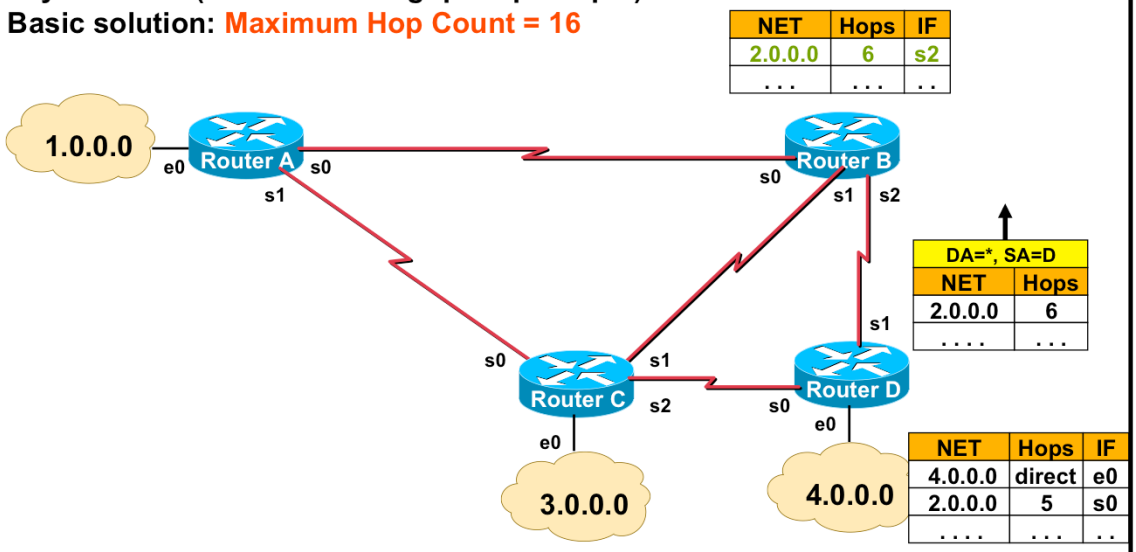
And a new connection established between router B and router D. Now, a normal routing update is send from router D to router B (with information about net 2, of course).



Router B doesn't know where network 2 is gone. So he sends information about network 2 (increasing hop count by 1) to every neighbor router.

Count To Infinity (4)

Count to Infinity situations cannot be avoided in any situation (drawback of signpost principle)
Basic solution: Maximum Hop Count = 16



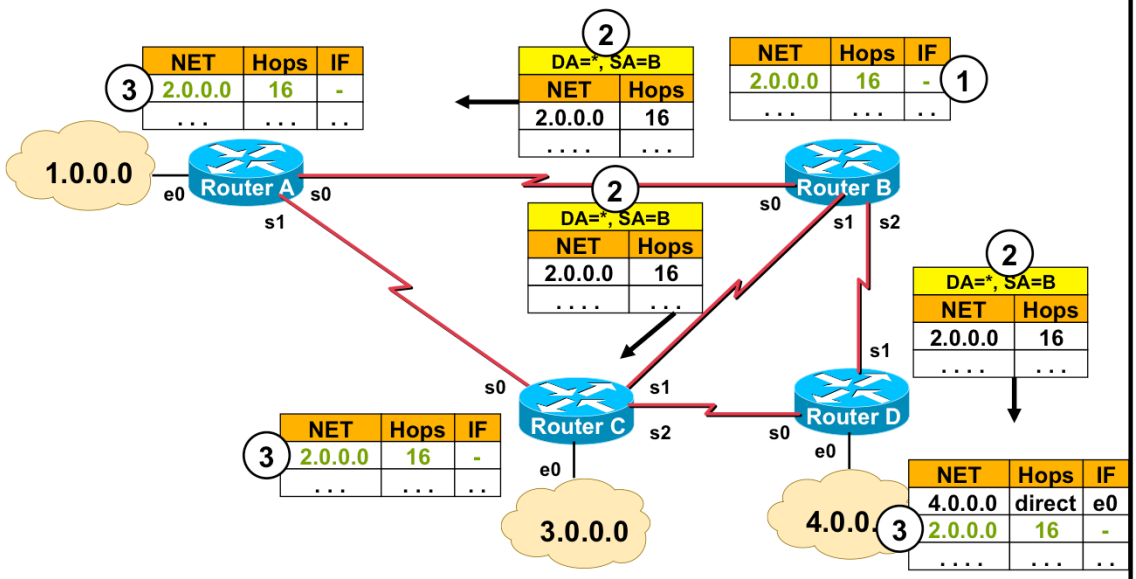
 IP Routing, v6.2 67

Count to infinity accurse. Only a “Maximum Hop Count”, the basic solution to avoid count-to-infinity in a distance vector network, can stop this problem.

Now we have only a hop count up to 16 instead of infinity. Of course during that time a routing loop exists in the network.

Maximum Hop Count = 16

Reaching hop count 16, the route is marked as INVALID (1) and propagated for a certain time (2) to inform neighbors about unreachablity of network 2.0.0.0.



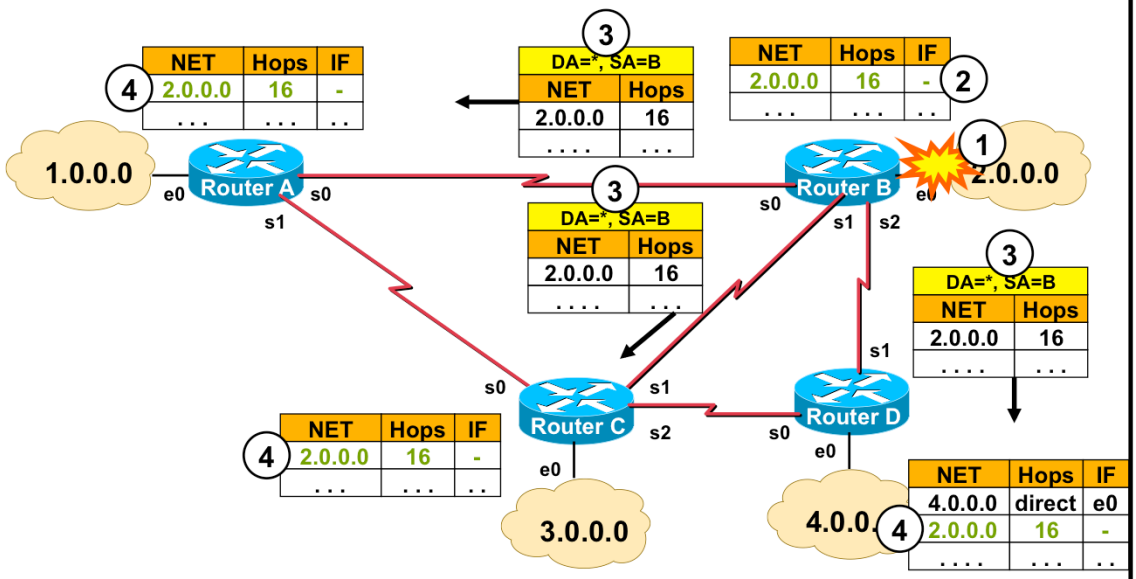
After 16 Hops the Net 2 is now marked as invalid and this is again propagated as usual to the neighboring routers.

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Maximum Hop Count = 16

Upon network failure, the route is marked as INVALID (hop count 16) and propagated.



Having defined maximum hop count, router B has now also means for immediately telling unreachability of network 2 after detecting the break of interface e0. Announcement of hop count 16 for a network means that this network is not reachable any more and should be removed from the routing table. Count-to-infinity (or better count to 16) will not happen any more and hence no routing loop will occur any longer in the network in such a situation. Of course, the unreachabilty-information would be propagated deeper into the network if there are additional routers.

Maximum Hop Count

- Defining a maximum hop count of 16 provides a basic safety factor
- But restricts the maximum network diameter
- Routing loops might still exist during 480 seconds (16×30s)
- Therefore several additional measures are necessary
  - Split Horizon
  - Poisen Reverse
  - Hold Down
  - Triggered Update

The maximum hop count is a basic safety factor, but it is also the main drawback of RIP. It restrict the maximum network diameter, and the rooting loops exist for 480 seconds. During Count to Infinity there is a bad routing and the network must deal with unnecessary traffic. So we need other measures like Hold down or Poison Reverse.

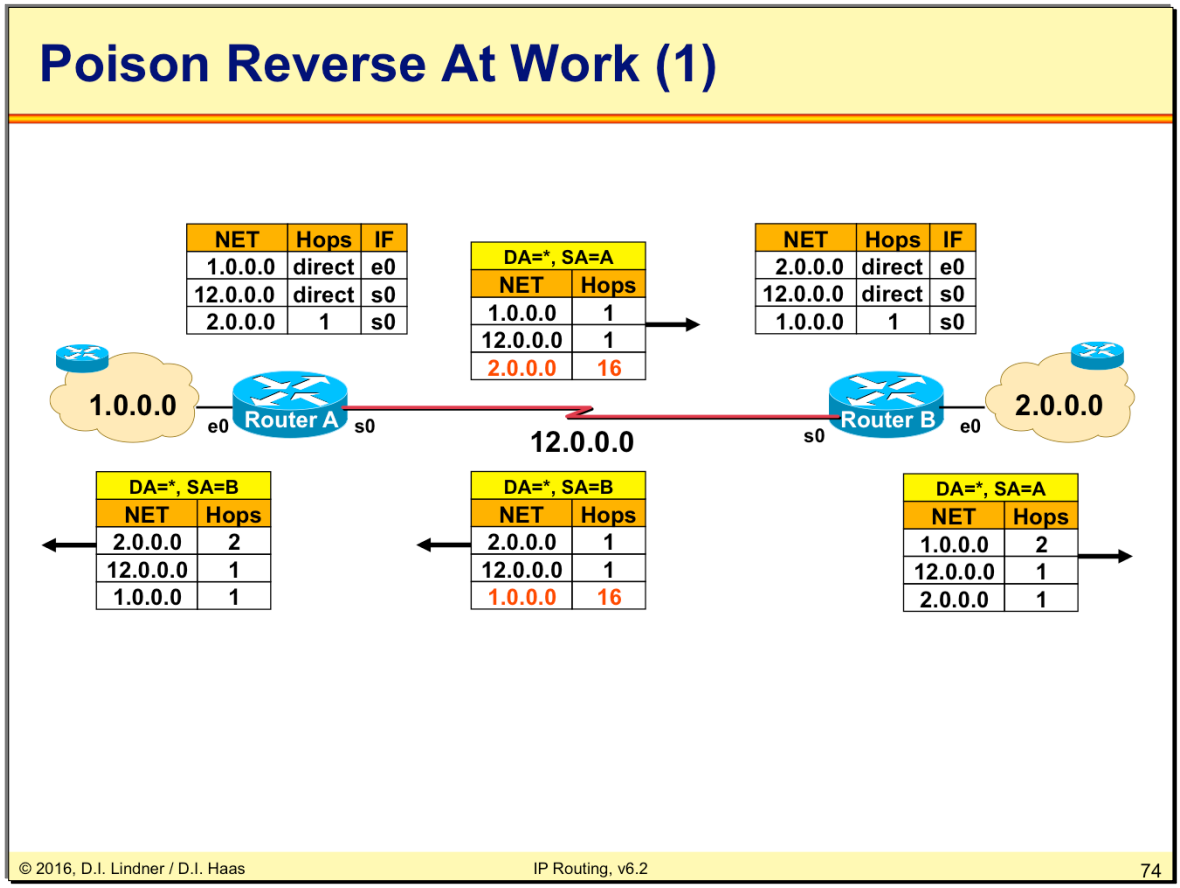
Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Additional Measures (1)

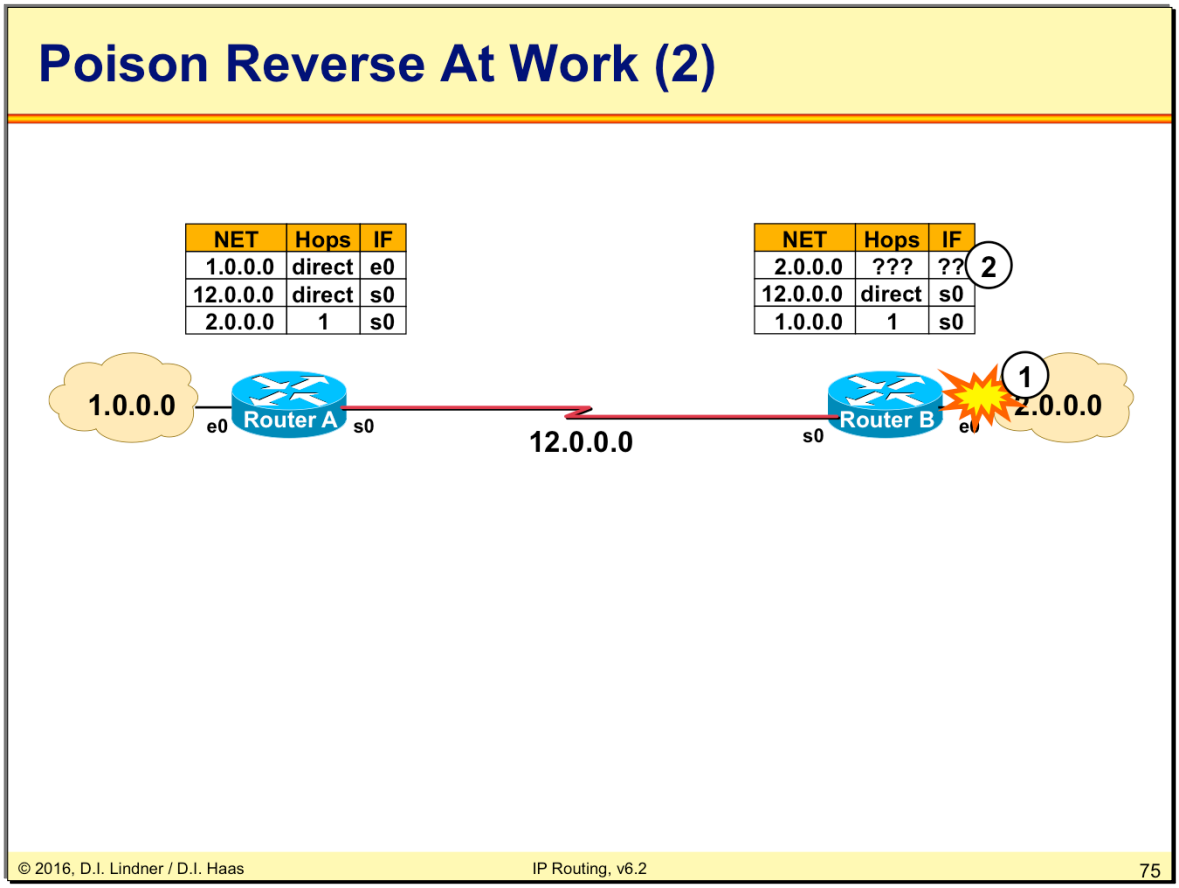
- **Split Horizon**
  - Suppressing information that the other side should know better
    - Used during normal operation but cannot prevent routing loops !!!
  - Remember: good news overwrite bad news
    - Unreachable information could be overwritten by uninformed routers (which are beyond scope of split horizon)
- **Poison Reverse**
  - Alternate approach split horizon
  - Declare learned routes as unreachable
  - "Bad news is better than no news at all"
  - Stops potential loops due to corrupted routing updates

Today's RIP implementations either use split horizon or poison reverse. They are compatible to each other.

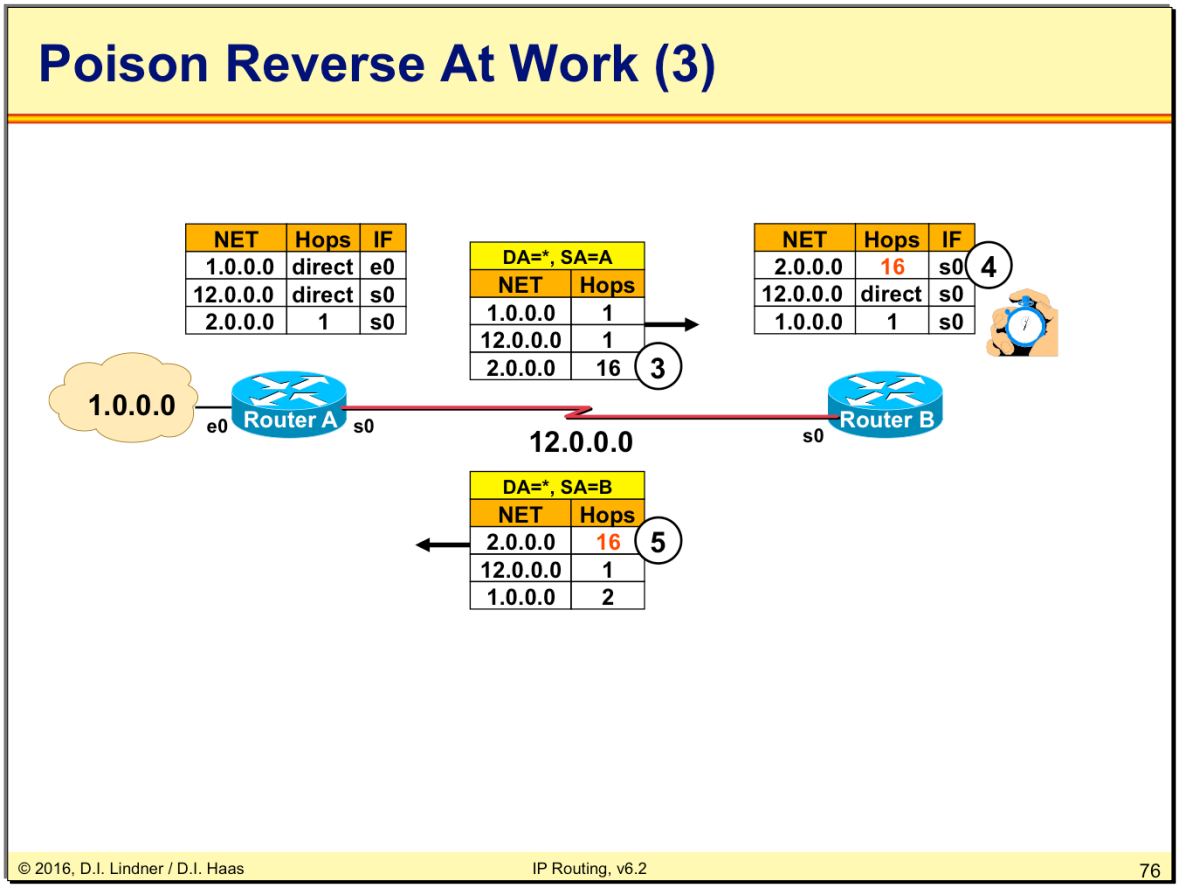


Poison reverse includes also reverse routes in updates, but sets their metrics to infinity. This is safer than simple split horizon: If two gateways have routes pointing at each other, advertising reverse routes with a metric of 16 will break the loop immediately.

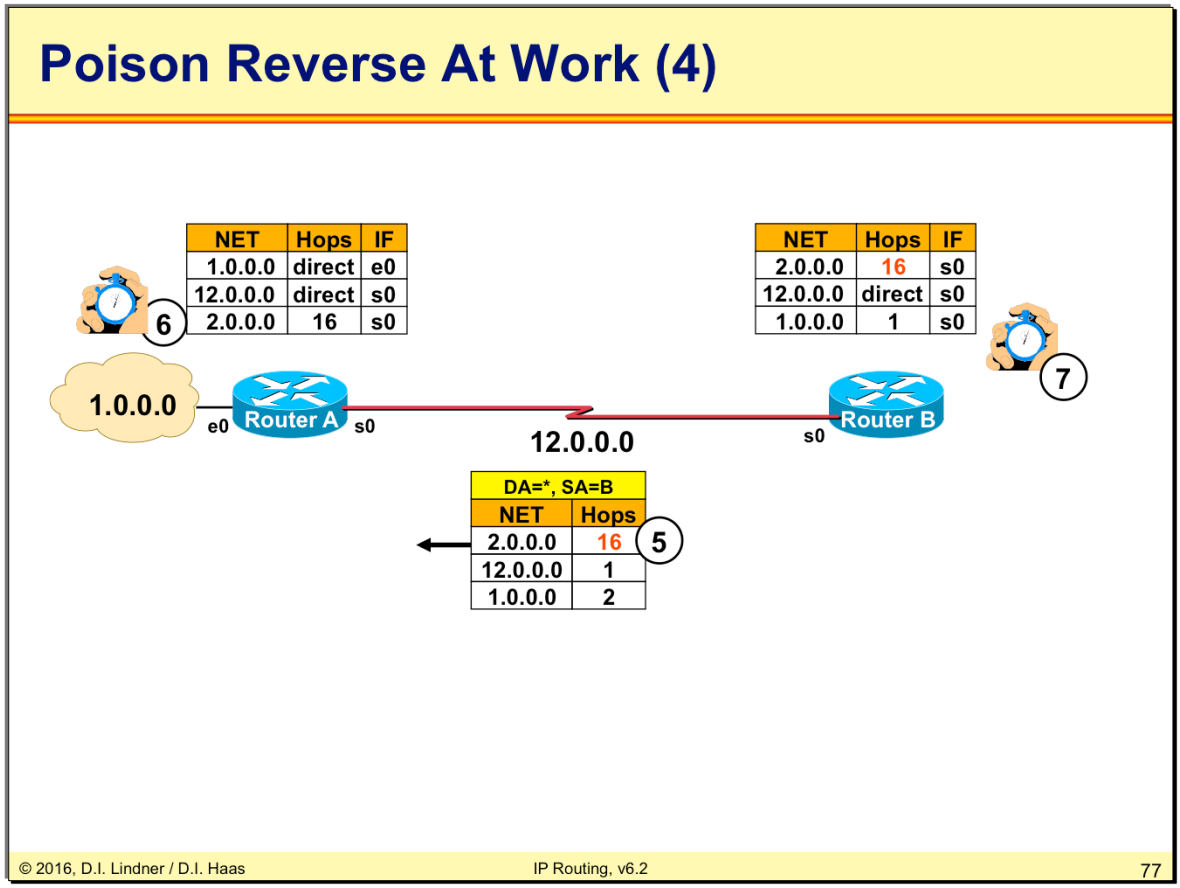
Note: Poison reverse is not used by Cisco Routers (however poison updates are indeed used when e. g. an interface goes down).



In this example we see what would happen if network 2 crashes (1). Immediately, router B has no more information about this net (2). What would happen if router A sends a routing update now?



Now router B receives a routing update from router A including unreachability information about network 2 (3). Because router B has no information about network 2 he adapts this unreachability information in his routing table (4) and sending his normal routing updates to router A, hereby increasing the hop count to 16 (5). In router B the unreachability information about network 2 will now times out after 180 seconds.



Upon receiving update

from router B (5) now router A increases the hop count to 16 (6). In router A the unreachability information about network 2 will now time out after 180 seconds.

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Additional Measures (2)

- **Hold Down**
  - Guarantees propagation of bad news throughout the network
  - Routers in hold down state ignore good news for 180 seconds
  - Basic idea:
    - Network-failure message requires a specific amount of time to spread across the whole network (like a wave)
    - With Hold Down, all routers get the chance to receive the network-failure message
    - Inconsistent routing-tables and rooting-loops are avoided

RIP needs long time to send bad news over the whole network (remember the 480 seconds). To guarantee that the bad news can be sent throughout the network, the hold down measure is implemented. After a router receives “bad news” he will ignore all “good news” about the same route for 180 seconds.

Note: Hold-down timers are not explicitly required by RFC 1058. However most vendors (also Cisco) implemented it.

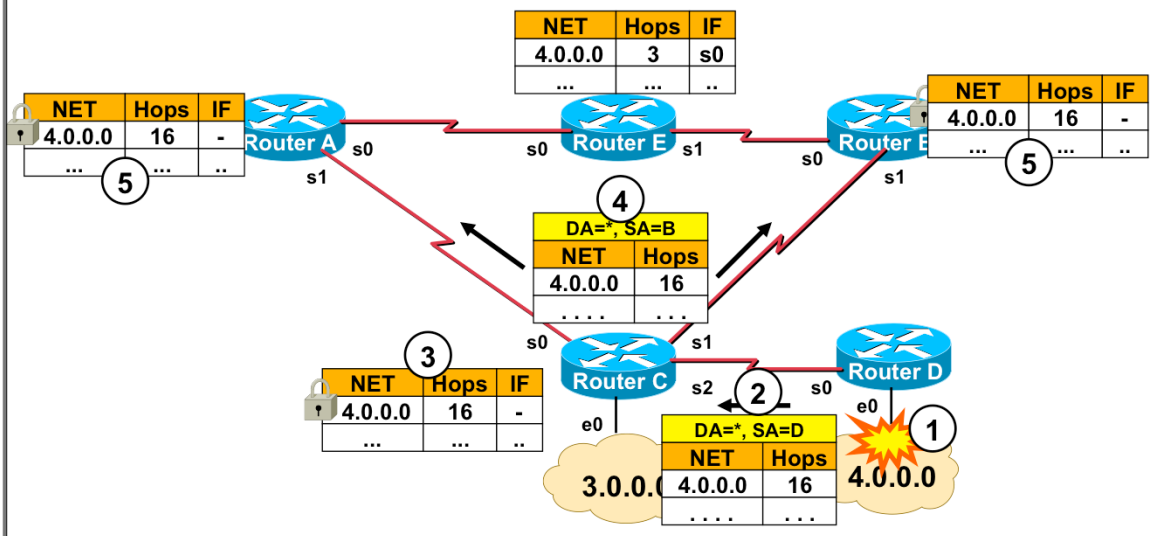
Although split Horizon is a good means to avoid temporary routing-loops and to improve the convergence time in simple network topologies, in complex network topologies require an additional tool to avoid temporary rooting-loops: Hold Down.

If a router gets information about a network failure, it ignores further information about that network from other routers for a specific duration of time (typically 240) seconds.

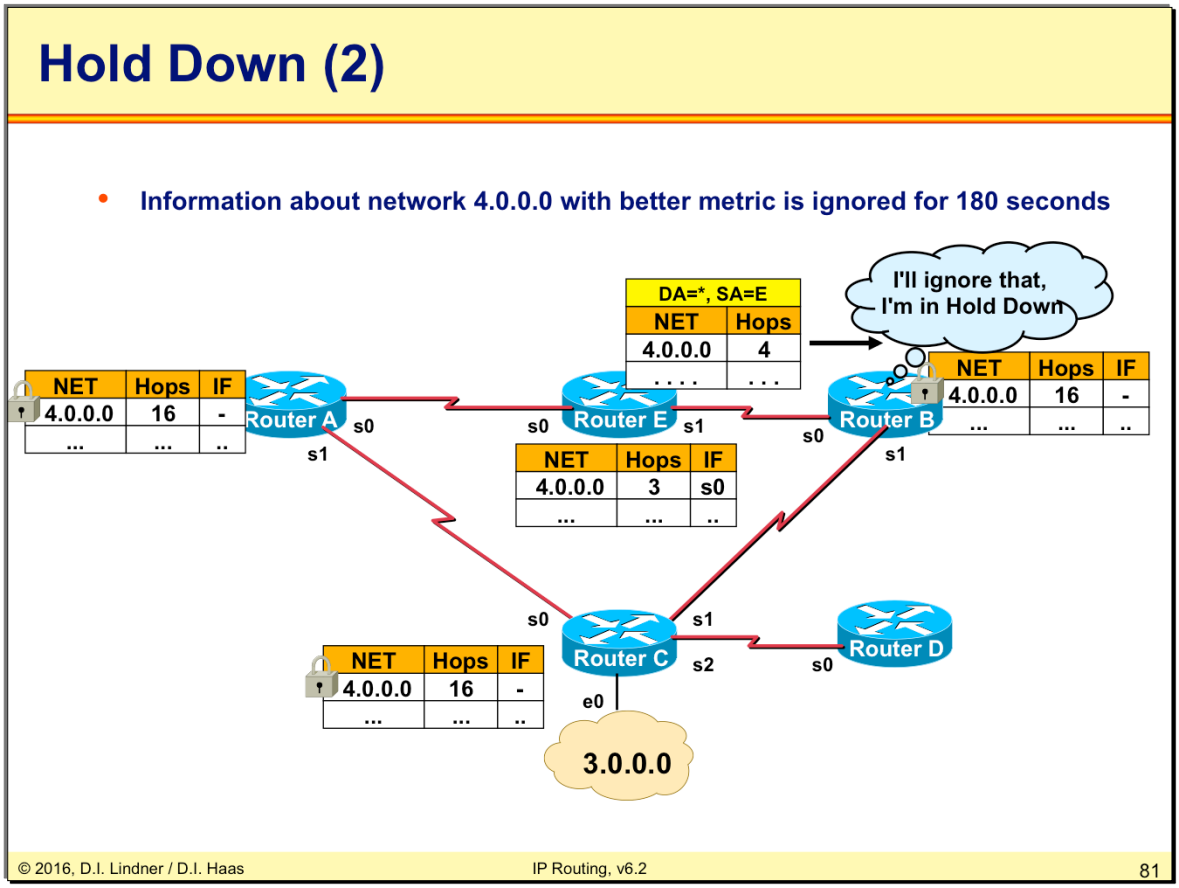
Disadvantages of Hold Down: Can lead to longer convergence time as maybe necessary because even alternate paths are not used during hold down time. The only event which stops the hold down before the timer expires is when the router receives an update about the network from that router which has forced him into hold down.

Hold Down (1)

- Router C receives unreachable message (4.0.0.0, 16) from router D
- Router C declares 4.0.0.0 as invalid (16) and enters hold-down state



In this example we see the functionary of Hold Down. After Net 4 crashes (1), router D send this information to Router C (2). Router C added this information and activate “hold down” (3). After this he sends this information to his neighbor routers (4), which do the same after they receive the information about net 4 (5).



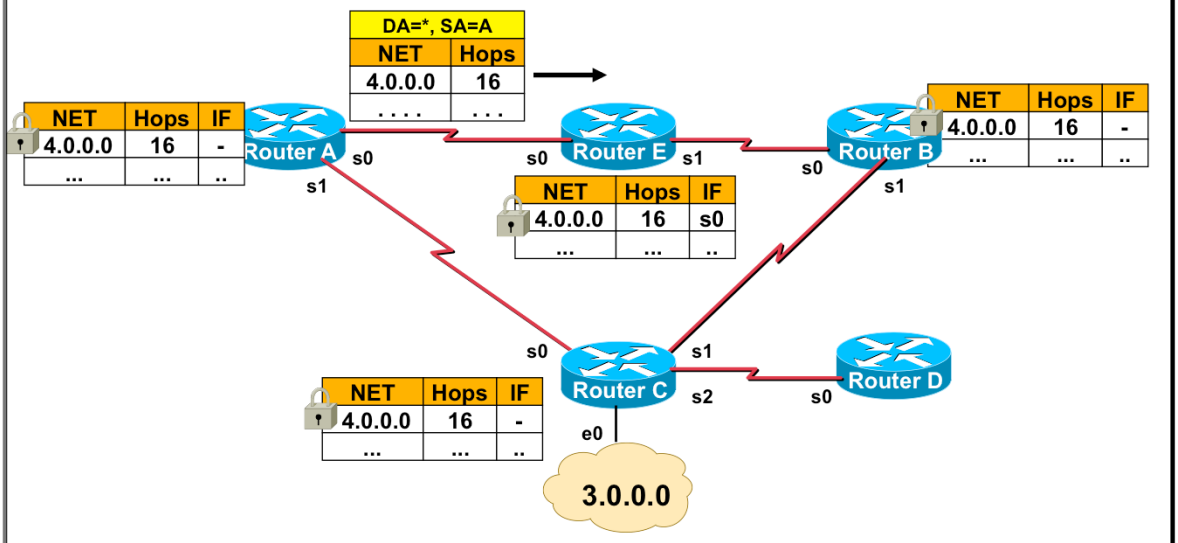
Router E didn't get information that net 4 crashes yet, so he normally sends his routing update.

Recognize that although split horizon is turned on router E will sent the information about network 4.0.0.0 / 4 hops out on interface s1 because router E has learnt about network 4.0.0.0 by a routing update from router A seen on interface s0, hence split horizon rule is obeyed in such a case.

But the information from router E couldn't overwrite routing information of router B or router A. Because these router are in the "hold down" status, and ignore these update messages.

Hold Down (3)

- Time enough to propagate the unreachabilty of network 4.0.0.0



Soon every router knows that network 4 is unreachable.

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary **FYI**
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Triggered Update / Timer Synchronization

- **To reduce convergence time, routing updates are sent immediately upon events (changes)**
  - New network connected to the router
  - Local network crashes
- **On receiving such a different routing update a router should also send immediately an update**
  - Called "**Triggered Update**"
- **In case of many routers on a single network**
  - Processing load might affect update timer
  - Router timers might get synchronized
  - Collisions will occur more often
- **Therefore either use**
  - External timer or add a small random time to the update timer (30 seconds + RIP_JITTER = 25...35 seconds)

 IP Routing, v6.2 84

To speed up the convergence time, "triggered update" has been introduced. After a router notice a network failure, he immediately sends a routing update to indicate this failure (hop-count =16). So the router didn't wait for the expiration of the 30 seconds. Triggered update can used with all events (e.g. a new link established) but triggered Update without employing additional methods (like Split Horizon) cannot avoid routing-loops for 100%.

RIP Timers Cisco

- **UPDATE (30 seconds)**
  - Period to send routing update
- **INVALID (180 seconds)**
  - Aging time before declaring a route invalid ("16") in the routing table
- **HOLDDOWN (180 seconds)**
  - After a route has been invalided, how long a router will wait before accepting an update with better metric
- **FLUSH (240 seconds)**
  - Time before a non-refreshed routing table entry is removed

The FLUSH timer is also known as "Garbage Collection Timer" and RFC 1058 suggests additional 120 seconds after expiring of the INVALID timer.

HOLDDOWN timers are not explicitly required by RFC 1058, however they are supported by most implementations today, e. g. by Cisco IOS. Note that the FLUSH timer expires before the HOLDDOWN timer.

RIP Disadvantages

- **Big routing traffic overhead**
  - Contains nearly entire routing table
  - WAN links (!)
- **Slow convergence**
- **Small network diameter**
- **No discontiguos subnetting**
- **Only equal-cost load balancing supported**
  - (if you are lucky)

RIP is an old protocol and only used in small networks.

Summary RIPv1

- **First important distance vector implementation (not only for IP)**
- **Main problem: Count to infinity**
  - Maximum Hop Count
  - Split Horizon
  - Poison Reverse
  - Hold Down
- **Classful, Slow, Simple**

Agenda

- **Introduction to IP Routing**
- **RIP**
  - Introduction
  - Split Horizon
  - Count-To-Infinity
  - Max-Hop-Count
  - Poison Reverse
  - Hold Down
  - Some Details and Summary
  - RIP Version2
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**

Why RIPv2?

- **Need for Subnet information and VLSM**
- **Need for Multicast Routing Updates**
  - RIPv1 used DA=255.255.255.255
    - Seen by each IP host
    - Slows down other IP stations
  - RIPv2 uses DA=224.0.0.9
    - Only RIPv2 routers will receive it
- **Need for Next Hop Addresses for each route**
- **Need for External Route Tags**

 IP Routing, v6.2 89

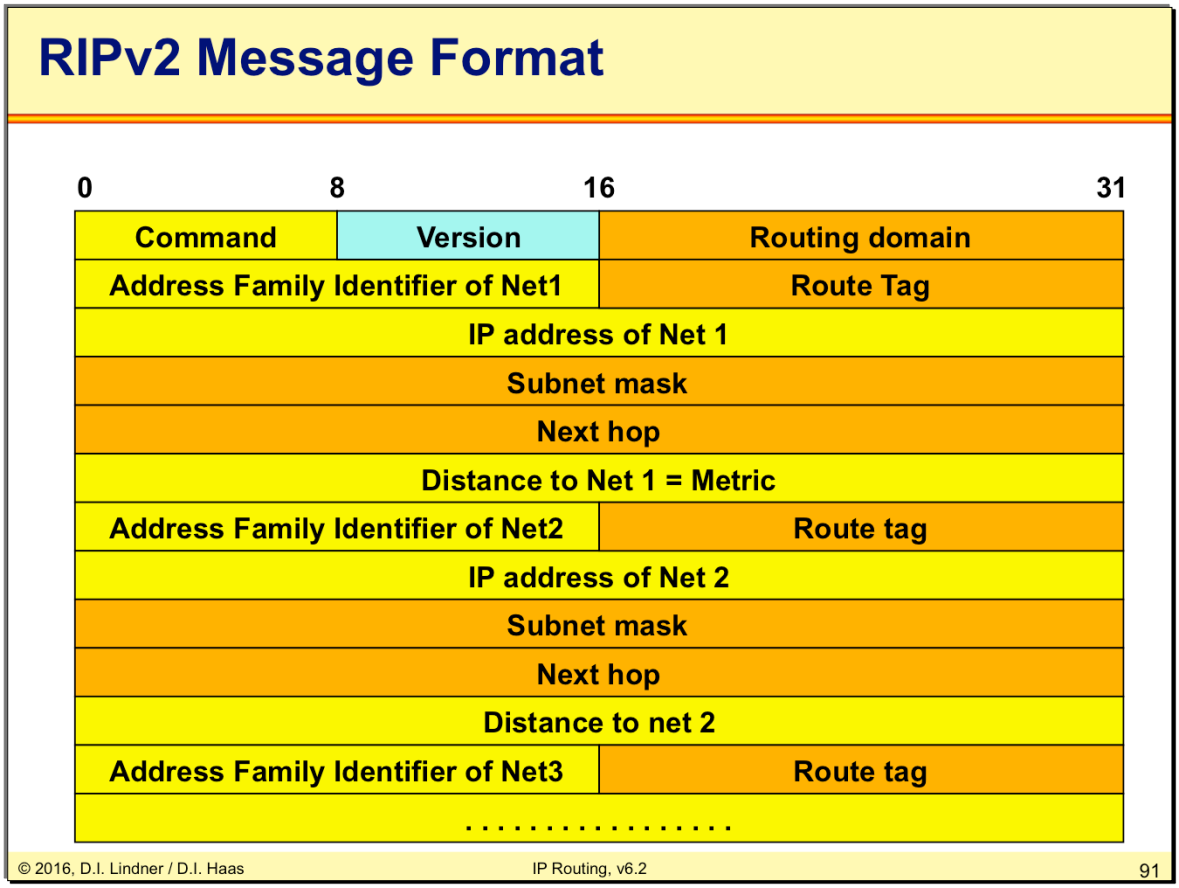
Because Subnetting and VLSM get more important RIPv2 was created. RIPv2 was introduced in RFC 1388, "RIP Version 2 Carrying Additional Information", January 1993. This RFC was obsoleted in 1994 by RFC 1723 and finally RFC 2453 is the final document about RIPv2.

In comparison with RIPv1 the new RIPv2 also support several new features such as, routing domains, route advertisements via EGP – protocols or authentication.

RIPv2 uses the IP-Address 224.0.0.9 to transfer his routing updates. With this advantage only RIPv2 routers see this messages, and will not slow down different stations as done with RIPv1 and IP-limited broadcast addresses.

RIPv2

- **RFC 2453 specifies a new, extended RIP version:**
  - RIPv2 is RFC category “Standard”
  - RIPv1 is RFC category “Historic”
- **RIPv2 is an alternative choice to OSPF**
  - OSPF has the touch to be more complicated!
- **Several new features are supported:**
  - Transmission of subnet-masks
  - Transmission of next hop redirect information
  - Routing domains and route tags
  - Route advertisements via EGP - protocols
  - Authentication
- **RIPv2 is a classless routing protocol**



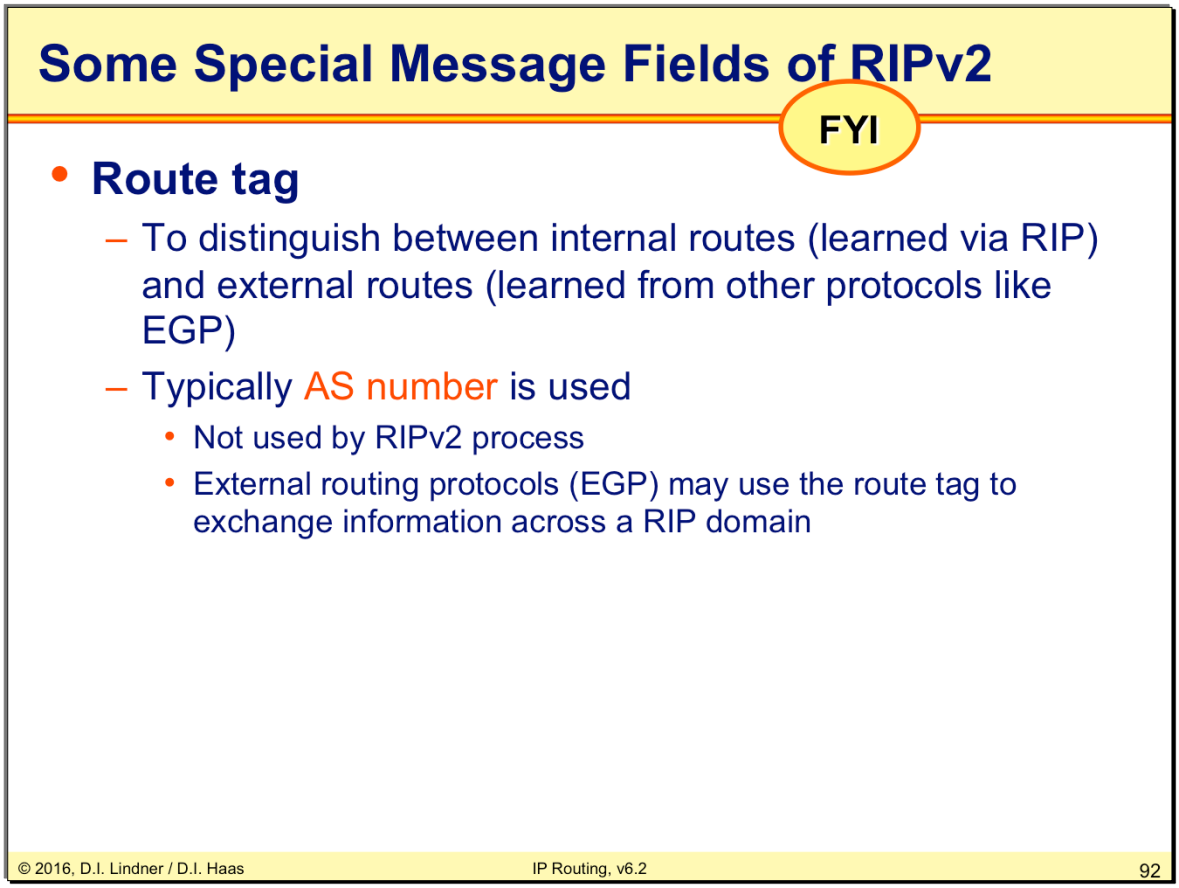
RIPv2 utilizes the unused fields of the RIPv1 message-format. New fields are the “routing tag”, “subnet mask” and the “next hop”.

RIPv1 used version "1“, RIPv2 uses version "2" (*surprise*).

According RFC the next two bytes are unused. However, some implementations carry the routing domain here which is simply a process number. The routing domain indicates the routing-process for which the routing-update is destined. Now routers can support several domains within the same subnet.

AFI , metric and command fields have the same meaning as for RIPv1.

Subnet mask contains the subnet-mask to the "IP address"-field. Now discontigous subnetting and variable length subnet masks (VLSM) techniques are supported.



For example if external routes are learnt by EGP and need to be redistributed from EGP into RIPv2, these routes can be tagged. So the other RIPv2 routers know which networks are internal and which are external. Filterinmg and policing function may be applied on routes depending on the route tag. E.g. internal routes should not be advertized to the outside world but external routes should be further propagated to other autonomous systems.

Some Special Message Fields of RIPv2

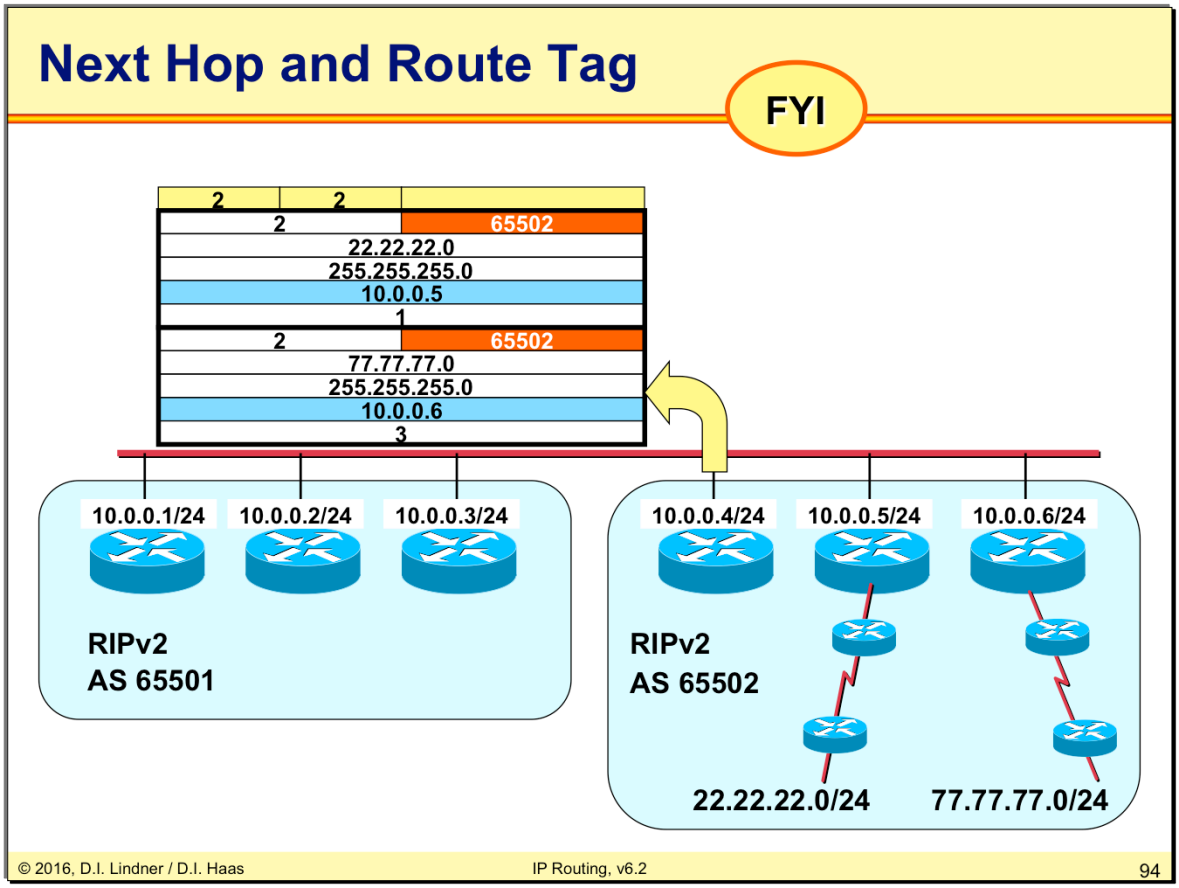
FYI

- **Next hop**
  - Datagrams for the network specified in the "IP address" - field have to be redirected to that router whose IP address is specified in the "next hop" field
    - This next-hop router must be located in the same subnet as the sender of the routing-update
    - A next hop value of 0.0.0.0 indicates, that the sender-router acts as next hop itself for the given network
  - Identifies a better next hop address than implicitly given by SA of the announcing router
  - Especially useful on broadcast multi-access network for peering
    - Indirect routing on a broadcast segment would be ...silly.

 IP Routing, v6.2 93

With the „next hop“ router announces which networks can be reached over other routers.

Note that the next-hop router must be located in the same subnet as the sender of the routing-update.



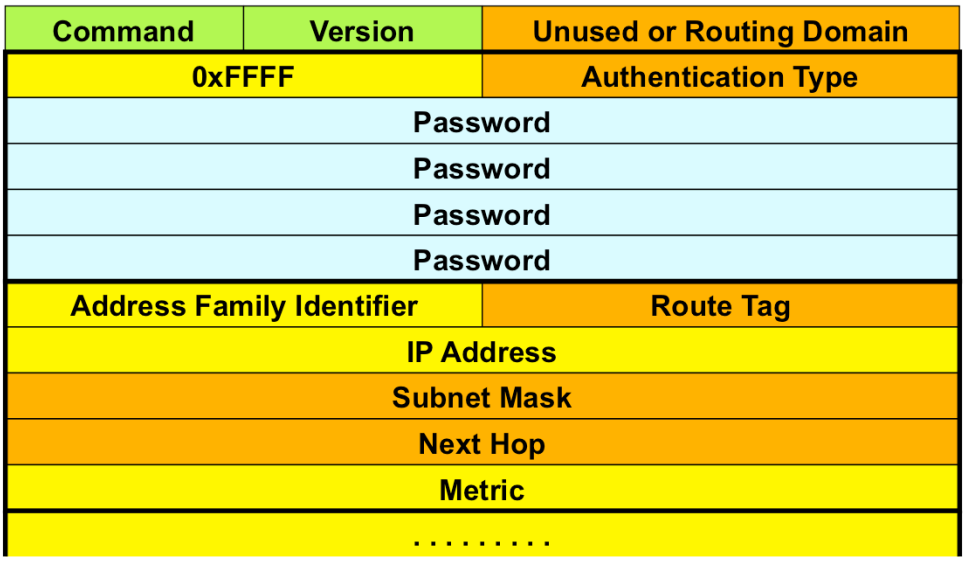
In the picture above there are two different autonomous systems on the same LAN. The routers in the first AS use RIPv2 and in the second AS use BGP. Each entry assigned a AS number (65501/65502). The Left AS could apply policies on these special (external) routes or redistribute them with BGP to some other ASs. Note that only 10.0.0.4 speaks RIPv2, so for efficiency only this one advertises the external routes (22.22.22.0/77.77.77.0) but by indicating the true next hops. This is an important special rule on shared medium (true next hops must be indicated) !

Authentication

- **Hackers might send invalid routing updates**
- **RIPv2 introduces password protection as authentication**
- **Initially only 16 plaintext characters (!)**
  - Authentication type 2
- **RFC 2082 proposes keyed MD-5 authentication**
  - Authentication type 2
  - Multiple keys can be defined
  - Updates contain a key-id and an unsigned 32 bit sequence number to prevent replay attacks
- **Cisco IOS supports**
  - MD5 authentication (Type 3, 128 bit hash)

IF a router receives routing updates without valid authentication are ignored by the receiving router, because only trusted router are accepted.

Authentication



Up to 24 route entries

 IP Routing, v6.2 96

The picture above shows a RIPv2 Message which contains an password authentication entry. The password is only a plain text. If the password is under 16 octets, it must be left-justified and padded to the right with nulls.

Address Family Identifier = hex FFFF

If this value is seen in the first AFI of net entry then authentication is used for that routing update.

Authentication Type: tells which kind of authentication is used and also which format the authentication data has.

Type 2 indicates "Password"

Type 3 is “Keyed Message Digest Algorithm MD5”

Type 1 indicates "IP Route" (and is used in the MD5 trailer = last routing entry)

When using MD5 authentication, the first but also the last routing entry space is used for authentication purposes. The MD5 hash is calculated using the routing update plus a password. Thus, authentication and message integrity is provided.

Key Chain

FYI

- **Cisco's implementation offers key chains**
  - Multiple keys (MD5 or plaintext)
  - Each key is assigned a lifetime (date, time and duration)
- **Can be used for migration**
  - Key management should rely on Network Time Protocol (NTP)

Several independent routing domains running RIPv2 with different process numbers ("routing domain"). With using key chains this domains can be work together (synchronize) at a special time or date.

RIPv1 Inheritance

- **All timers are the same**
  - UPDATE
  - INVALID
  - HOLDDOWN
  - FLUSH
- **Same convergence protections**
  - Split Horizon
  - Poison Reverse
  - Hold Down
  - Maximum Hop Count (also 16 !!!)
- **Same UDP port 520**
  - Also maximum 25 routes per update
  - Equally 512 Byte payloads

 IP Routing, v6.2 98

RIPv1 uses many timers to regulate its performance. This timers are the same in RIPv2. The routing update timer is set to 30 seconds, with a small random amount of time added whenever the timer is reset. A route is declared invalid without being refreshed by routing updates during 90 seconds. The “holddown” status retains 180 seconds. In this time a router ignore update messages about a special network. After 240 Seconds (Flush timer) a non-refreshed routing table entry will be removed.

RIPv2 also using the same convergence protections such as Split Horizon, Hold Down, etc. Note that the Maximum Hop Count is still 16 to be backwards compatibility.

RIPv1 Compatibility

FYI

- **RIPv1 Compatibility Mode**
  - RIPv2 router uses broadcast addresses
  - RIPv1 routers will ignore header extensions
  - RIPv2 performs route summarization on address class boundaries
    - Disable: `(config-router)# no auto-summary`
- **RIPv1 Mode**
  - RIPv2 sends RIPv1 messages
- **RIPv2 Mode**
  - Send genuine RIPv2 messages

 IP Routing, v6.2 99

RIPv2 is totally backwards compatible with existing RIP implementations.
There is also an compatibility switch, which allows to chance between three different settings:
RIP-1 Modus. Only RIP-1 packets are sent
RIP-1 compatibility Modus. RIP-2 packets are broadcast
RIP-2 Modus. RIP-2 packets are multicast.
The recommended default for this switch is RIP-1 compatibility.

RIPv2 Summary

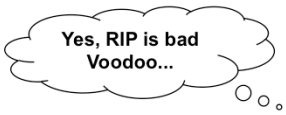
- **Most important: RIPv2 is classless**
  - Subnet masks are carried for each route
- **Multicasts and next hop field increase performance**
- **But still not powerful enough for large networks**

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Open Shortest Path First

- **Official (IETF) successor of RIP**
  - RIP is slow
  - RIP is unreliable
  - RIP produces too much routing traffic
  - RIP only allows 15 hop routes



- **OSPF is a link-state routing protocol**
  - "Open" means "not proprietary"
  - Inherently fast convergence
  - Designed for large networks
  - Designed to be reliable
- **OSPF's father: John Moy**
  - Version 1: RFC 1131
  - Version 2: RFC 2328 (244 pages !!!)
    - V2 first released in RFC 1583 obsoleted by RFC 2178

 IP Routing, v6.2 102

Distance vector protocols like RIP have several dramatic disadvantages. Examples are slow adaptation in case of network topology changes, size of routing update is proportional to network size and son on.

This led to the development of link-state protocols.

OSPF is the important implementation of link-state technique for IP routing.

OSPF was developed by IETF to replace RIP. In general link-state routing protocols have some advantages over distance vector, like faster convergence, support for lager networks.

Some other features of OSPF include the usage of areas, which makes possible a hierarchical network topologies, classless behavior, there are no such a problem like in RIP with discontigous subnets. OSPF also supports VLSM and authentication.

The Internet Engineering Task Force (IETF) strictly recommends to use OSPF for Interior Gateway routing (i. e. within an AS) instead of RIP or other protocols. Integrated IS-IS is an alternative routing protocol but not explicitly recommended by the IETF. Note that IS-IS has been standardized by the ISO world.

OSPF version 2 has been specified in RFC 2328. Note that there are a lots of additional RFCs around OSPF. Use http://www.rfc-editor.org/rfcsearch.html to find them all.

OSPF Base Principles

- Every router knows topology of the whole network including subnets and routers
  - "Roadmap"
- Topology (roadmap) stored in router's OSPF database
- Shortest Path First (SPF) algorithm applied to find the best path
  - Invented by E. W. Dijkstra
  - Creates a (loop-free) tree with local router as source
  - Is used to find the best path by calculating very efficiently all paths to all destinations at once; best path is entered into the routing table
- Changes are flooded over the network to update the OSPF database
  - Like traffic announcements used by car navigation systems
  - LSA (Link State Advertisements)

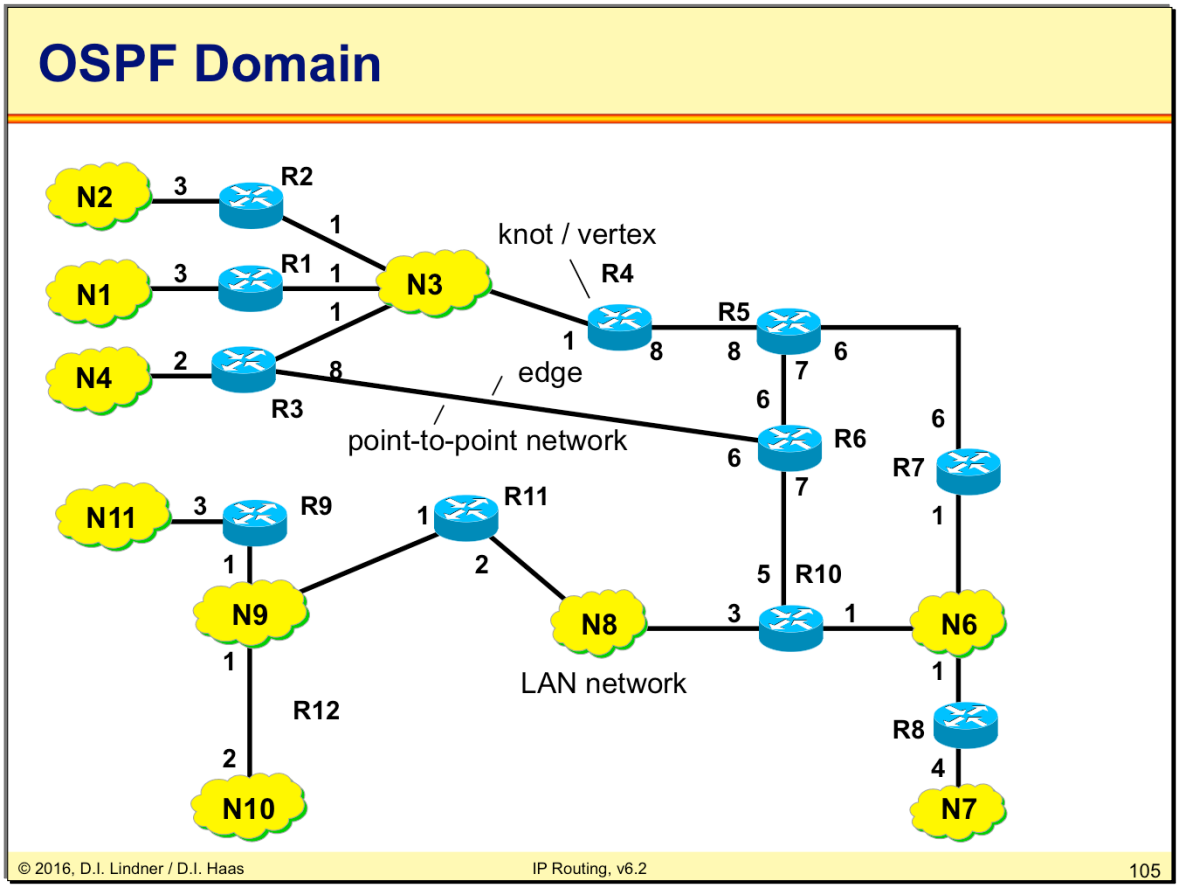
 IP Routing, v6.2 103

The Dijkstra's SPF algorithm is generally used in graph theory and was not invented especially for IP routing. The most interesting point on the SPF algorithm is its efficiency. SPF is capable to calculate all paths to all destinations at once. The result of the SPF algorithm is a loop-less tree with the local router as source.

Both (Integrated) IS-IS and OSPF use Dijkstra's famous Shortest Path First (SPF) algorithm to determine all best paths for a given topology.

OSPF Topology Database

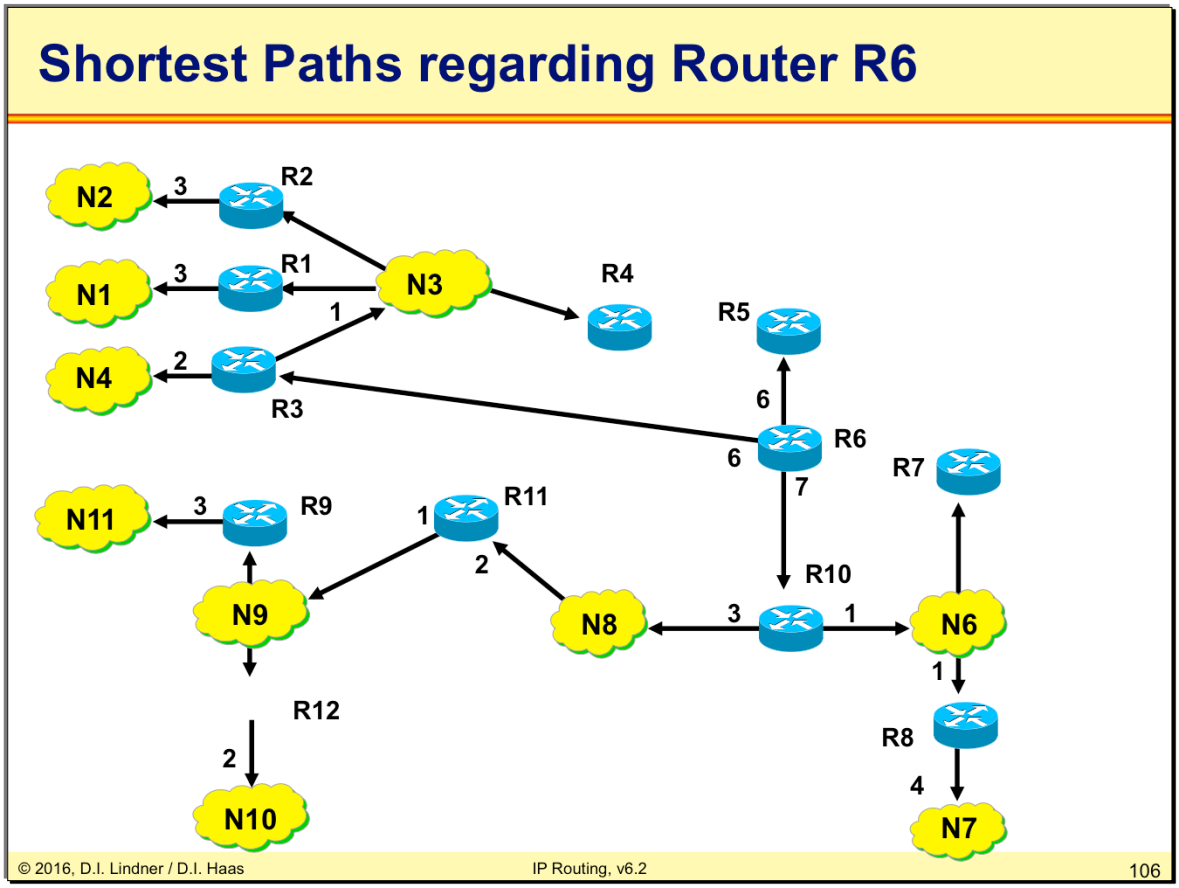
- **Every router maintains a topology database**
  - Like a "network roadmap"
  - Describes the whole network !!
    - Note: RIP provides only "signposts"
- **Database is based on a graph**
  - Where each knot (vertex) stands for a router
  - Where each edge stands for a subnet
    - Connecting the routers
    - Path-costs are assigned to the edges
- **Router uses the graph**
  - To calculate shortest paths to all subnets
    - Router itself is the root of the shortest path



With this topology-database a router can calculate the best path to all destination-networks by applying Dijkstra's SPF (Shortest Path First) algorithms.

The topology-database describes all other possible paths too. So in critical situations (failures) the router can independently calculate an alternative path.

There is no waiting for rumors of other routers anymore which was the reason for several RIP problems.



After calculating the shortest path the routing table is constructed by just adding next hop and summary metric taken from the shortest path tree for every network.

Routing Table Router 6

NET-ID	NEXT HOP	DISTANCE
N1	R3	10
N2	R3	10
N3	R3	7
N4	R3	8
N6	R10	8
N7	R10	12
N8	R10	10
N9	R10	11
N10	R10	13
N11	R10	14

OSPF Ideas

- **Metric: "Cost" = 10^8/BW (in bit/s)**
  - Therefore easily configurable per interface
- **OSPF routers exchange real topology information**
  - Stored in dedicated topology databases
- **Now routers have a "roadmap"**
  - Instead of signposts (RIP)
- **Incremental updates**
  - NO updates when there is NO topology change
- **Fast convergence**
  - Almost no routing traffic in absence of topology changes

In the Cisco IOS implementation starting with 11.2, the cost is calculated automatically by the simple formula 10,000,000/BW.

Here the bandwidth parameters on a routers interface are used, thus it is especially important to configure it on the serial interfaces.

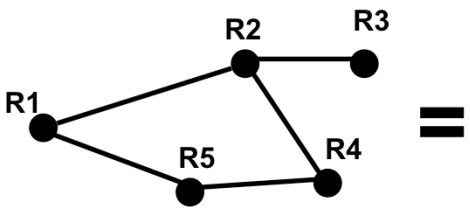
In other OSPF implementations cost must be configured manually for each of the interfaces.

OSPF—and other link state protocols—exchange true topology information which is stored in a dedicated database by each router. This database acts like a "roadmap" and allows a router to determine all best routes.

Note that once OSPF got the topology database there is no need to exchange further routing traffic—unless the topology changes. In this case only incremental updates are made.

What is Topology Information?

- The smallest topological unit is simply the information element ROUTER-LINK-ROUTER
- So the question is: Which router is linked to which other routers?
- Link-state



Link Database:

R1– R2
R1– R5
R2– R3
R2– R4
R4– R5

The Link Database exactly describes the roadmap

Obviously the dots are routers and the links between the routers are actually networks. The basic idea of OSPF and the topology table is that simple.

OSPF is actually much more complicated. There are 5 types of networks defined in OSPF: point-to-point networks, broadcast networks, non-broadcast multi-access networks, point-to-multipoint networks, and virtual links. Furthermore it is reasonable to divide the topology into multiple "areas" to increase performance ("divide and conquer"). These are the reasons why OSPF is a rather complex protocol. This is explained later.

OSPF Routing Updates / LSA

- **The routing updates are actually link state updates**
  - Parts of link state database are exchanged
  - Instead of parts of routing table (RIP)
  - **L**ink **S**tate **A**dvertisement (LSA)
- **Applying the SPF algorithm on the link state database**
  - Each router can create routing table entries by its own
- **LSAs are carried**
  - In small packets, forwarded by each router without much modifications through the whole OSPF domain (area)
  - Flooding principle
- **Much faster than RIP updates**
  - RIP must receive, examine, create, and send
- **Convergence time**
  - Detection time + LSA flooding + 5 seconds before computing the topology table = "a few seconds"

 IP Routing, v6.2 110

The Links State Updates LSUs are sent in a special packets – Link State Advertisements LSAs. There are several types of LSAs, depending on what kind of information is sent and which router originated it.

When the router gets a new information in its link state database it should send this information to all adjacent routers – flood. The packets are small, only the changes are sent and not the whole database. All other routers do the same, receive new information, update link state database, flood changes to others.

OSPF Areas – OSPF Performance

- **Large networks: "Divide and conquer" into areas**
  - LSA-procedures inside each area
  - But *distance-vector updates between areas*
- **Additional complexity because of performance optimizations**
  - Limit number of adjacencies in a multi-access network OSPF
  - Limit scope of flooding through "Areas"
  - Deal with stub areas efficiently
  - Learn external routes efficiently
  - Realized through different LSA types

 IP Routing, v6.2 111

Performance is very important with OSPF, to run SPF algorithm a CPU resources are required, to store a link state database an additional memory, compared to RIP we need much more router‘s resources. Some additional improvements were made to OSPF in order to improve performance. Areas were introduced to limit the flooding of LSAs, Stub Areas to minimize a link state database and routing tables.

Several types of LSAs were implemented:
Type 1 – Router LSA
Type 2 – Network LSA
Type 3 – Network Summary LSA
Type 4 – ASBR Summary LAS
Type 5 – AS External LSA
Type 6 – Group Membership LSA
Type 7 – NSSA External LSA
and others

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm **FYI**
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

About E. W. Dijkstra

- **Born in 1930 in Rotterdam**
- **Degrees in mathematics and theoretical physics from the University of Leyden and a Ph.D. in computing science from the University of Amsterdam**
  - Programmer at the Mathematisch Centrum, Amsterdam, 1952-62
  - Professor of mathematics, Eindhoven University of Technology, 1962-1984
  - Burroughs Corporation research fellow, 1973-1984
  - Schlumberger Centennial Chair in Computing Sciences at the University of Texas at Austin, 1984-1999
  - Retired as Professor Emeritus in 1999
  - 1972 recipient of the ACM Turing Award, often viewed as the Nobel Prize for computing
- **Died 6 August 2002**

Edsger W. Dijkstra (1930-2002)

Member of the Netherlands Royal Academy of Arts and Sciences, a member of the American Academy of Arts and Sciences, and a Distinguished Fellow of the British Computer Society. He received the 1974 AFIPS Harry Goode Award, the 1982 IEEE Computer Pioneer Award, and the 1989 ACM SIGCSE Award for Outstanding Contributions to Computer Science Education. Athens University of Economics awarded him an honorary doctorate in 2001. In 2002, the C&C Foundation of Japan recognized Dijkstra "for his pioneering contributions to the establishment of the scientific basis for computer software through creative research in basic software theory, algorithm theory, structured programming, and semaphores".

Dijkstra enriched the language of computing with many concepts and phrases, such as structured programming, separation of concerns, synchronization, deadly embrace, dining philosophers, weakest precondition, guarded command, the excluded miracle, and the famous "semaphores" for controlling computer processes. The Oxford English Dictionary cites his use of the words "vector" and "stack" in a computing context.

(Source: http://www.cs.utexas.edu)

"The question of whether computers can think is like the question of whether submarines can swim"

Edsger Wybe Dijkstra

Dijkstra's SP Algorithm

- **Famous paper "A note on two problems in connection with graphs" (1959)**
- **Single source SP problem in a directed graph**
- **Important applications include**
  - Network routing protocols (OSPF, IS-IS)
  - Traveler's route planner

Single source SP algorithms find all shortest paths to all vertices at once. The only difference to single-pair SP algorithms is the termination condition.

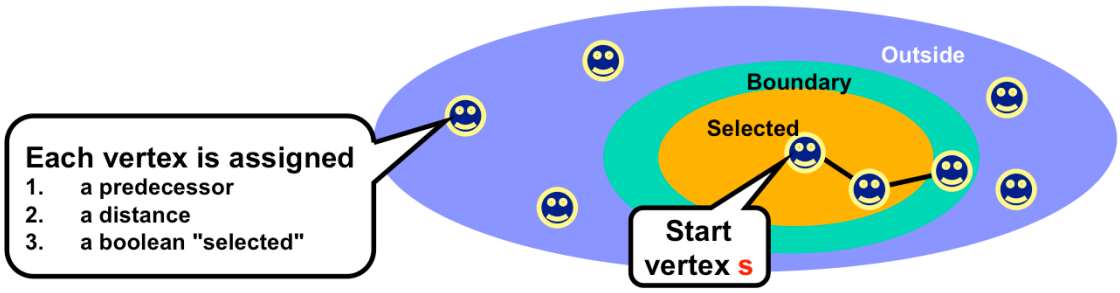
Terms

- **Graph G(V,E) consists of vertices V and edges E**
- **Edges are assigned costs c**
- **"Length" of graph c(G) = sum of all costs**
  - Assumed to be positive ("Distance Graph")
- **"Distance" between two vertices d(v,v') = min{c(p)}, p… path**
  - Can be infinite
- **p with c(p) = d(v,v') is called shortest path sp(v,v')**

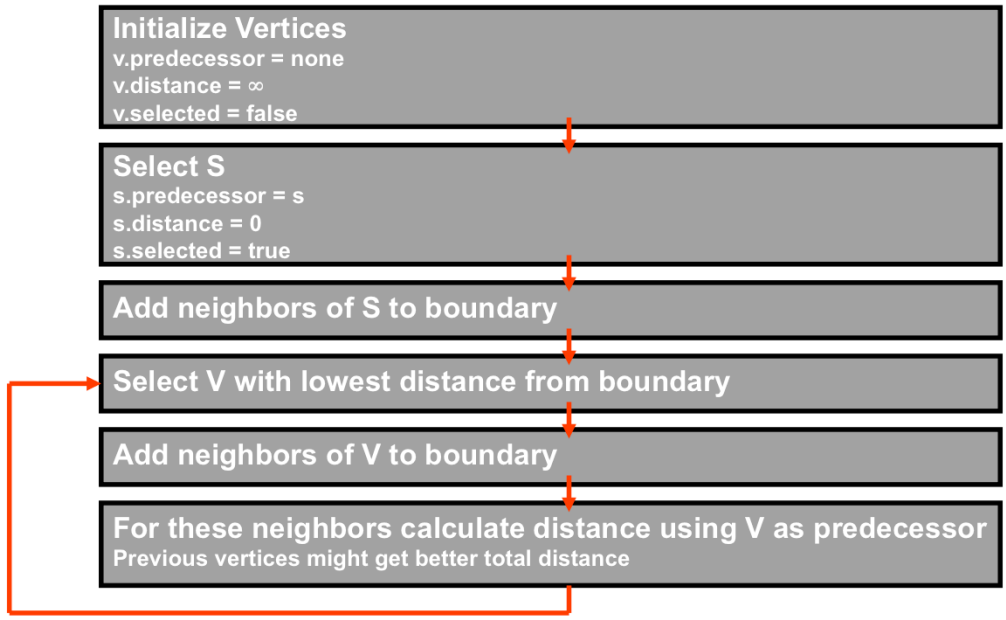
SPs are easier to calculate for distance graphs where the costs are only positive.

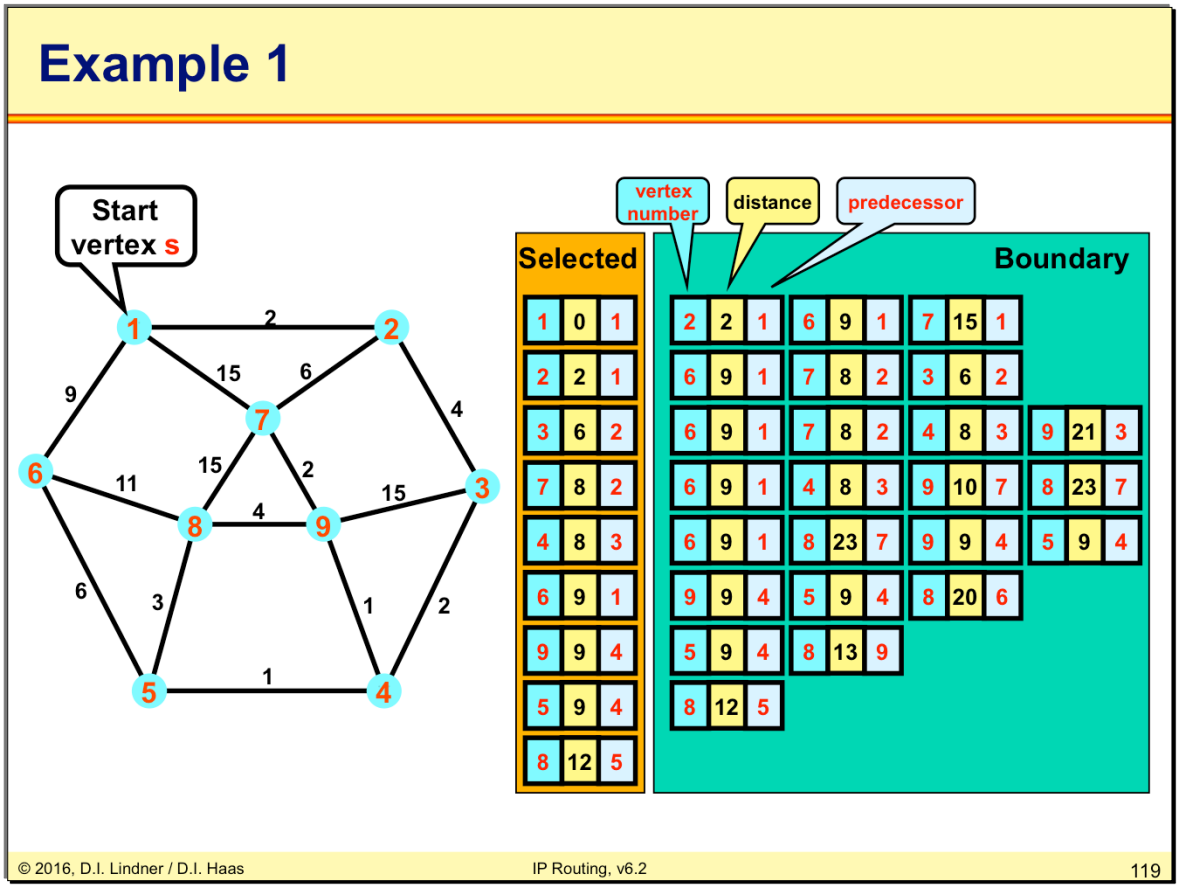
Definitions

- **Select start vertex s**
- **Three sets of vertices:**
  - Selected (sp already calculated)
  - Boundary (currently subject of calculation)
  - Outside (not yet examined)



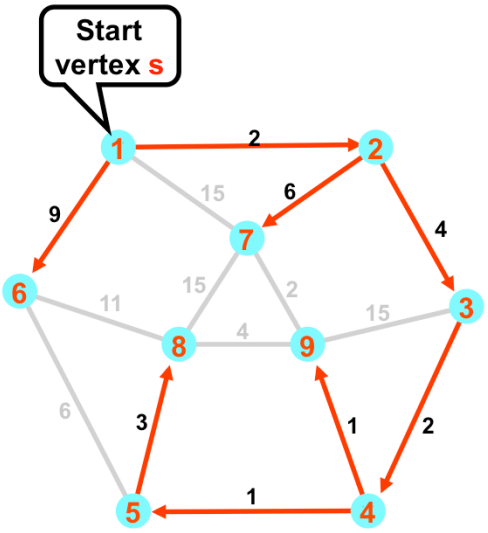
The Algorithm





Note that the left list ("Selected") is sometimes called the PATH list, and the right list ("Boundary") is sometimes called the TENT list (from tentative).

Result



Selected		
1	0	1
2	2	1
3	6	2
7	8	2
4	8	3
6	9	1
9	9	4
5	9	4
8	12	5

- Single source SP
- Minimal length
- Complete

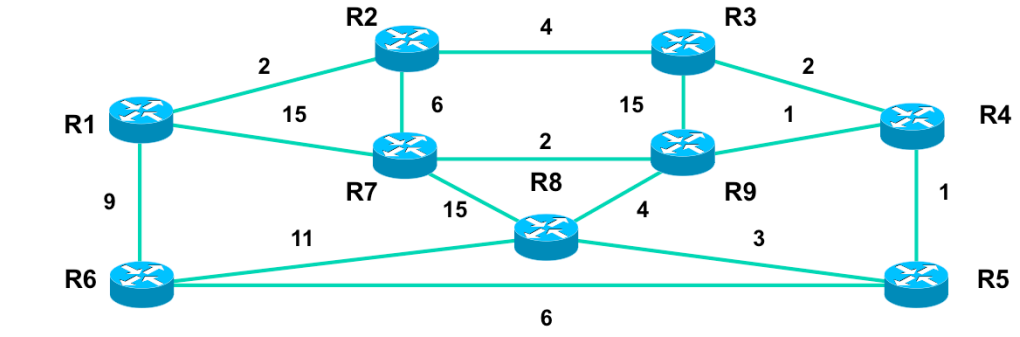
Performance

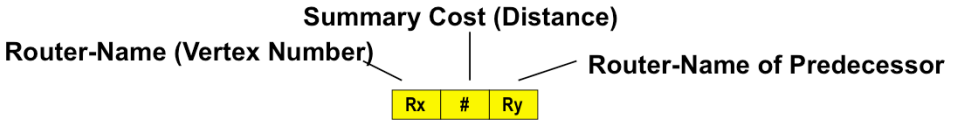
- **Greedy algorithm**
- **Most critical: Implementation of boundary data structure**
  - No explicit structure: $O(|V|^2)$
  - Fibonacci heap: $O(|E|+|V| \log |V|)$
- **Alternatives**
  - Bellman-Ford (RIP) algorithm
  - Floyd-Warshall algorithm
  - A* algorithm
    - Extends SPF with a estimation function to enhance performance in certain situations

The SPF algorithm is of “greedy” type. Dijkstra originally proposed to treat the boundary vertices like outside vertices, therefore no explicit data structure is needed for the boundary vertices. This implementation is efficient for graphs with lots of edges but not efficient with so-called "thin" graphs. One of the best implementations use Fibonacci heaps for boundary representation.

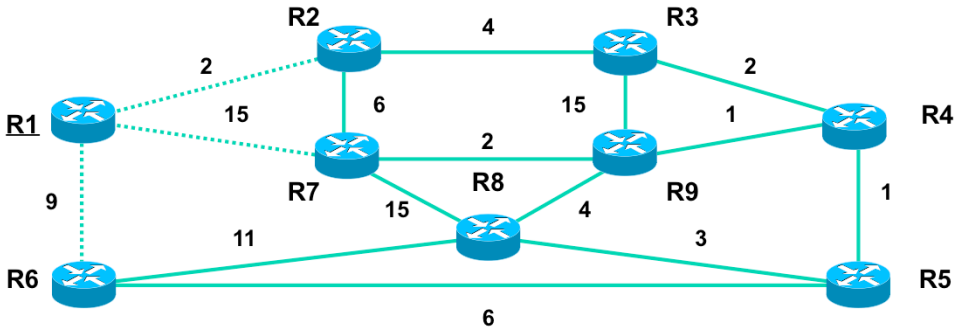
Alternative algorithms are for example the Bellman-Ford or the Floyd-Warshall algorithm, which bases on Bellman’s optimization principle (“if the shortest path from A to C runs over B, then the partial path AB must also be the shortest possible”).

Example 2 for Dijkstra Algorithm in Action





Select root (R1)



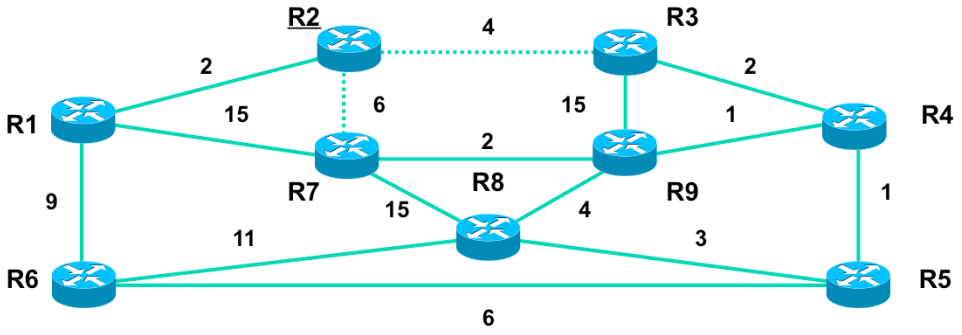
Selected

R1	0	R1

Boundary

R2	2	R1	R6	9	R1	R7	15	R1

Select router with lowest cost in boundary (R2), calculate cost for neighbours R3, R7



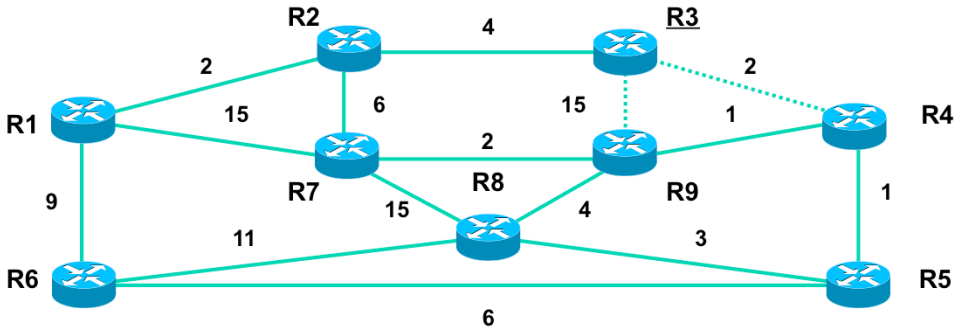
Selected

R1	0	R1
R2	2	R1

Boundary

R2	2	R1	R6	9	R1	R7	15	R1
R6	9	R1	R7	8	R2	R3	6	R2

Select router with lowest cost in boundary (R3), calculate cost for neighbours R9, R4



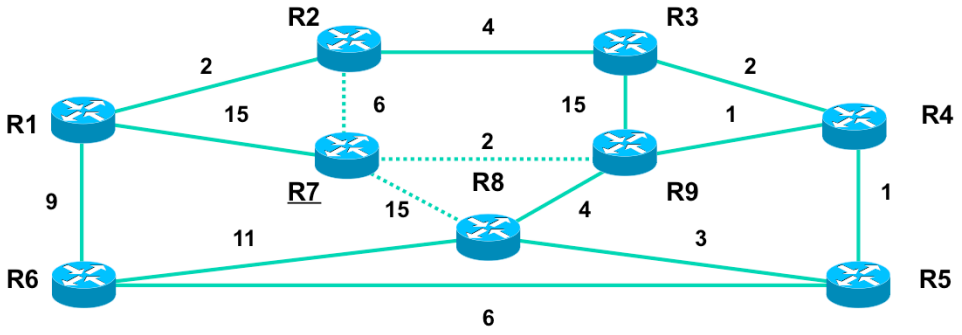
Selected

R1	0	R1
R2	2	R1
R3	6	R2

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3

Select one router with lowest cost in boundary (R7), calculate cost for neighbours R8, R9



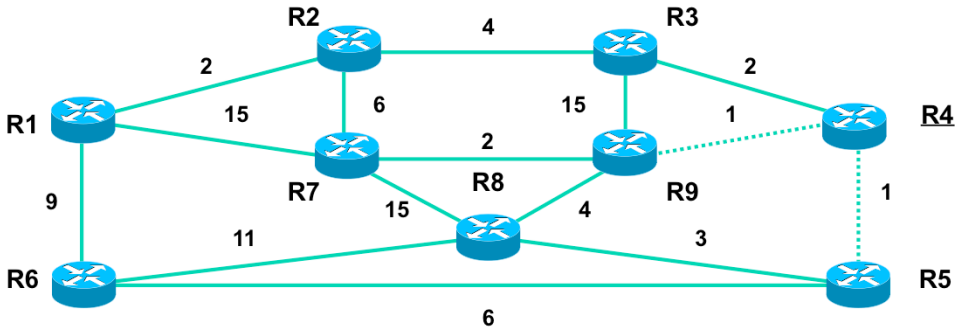
Selected

R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7

Select router with lowest cost in boundary (R4), calculate cost for neighbours R9, R5



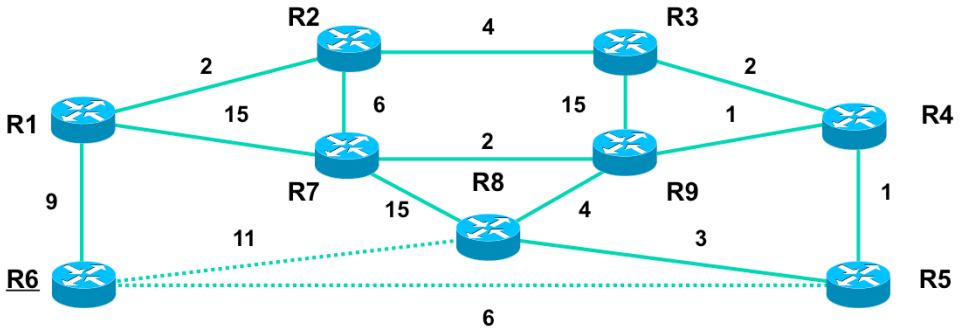
Selected

R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4

Select one router with lowest cost in boundary (R6), calculate cost for neighbours R5 and R8



Selected

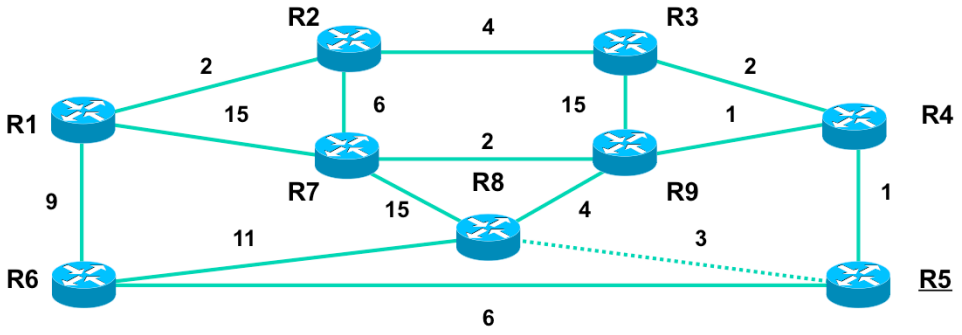
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			

 IP Routing, v6.2 128

Select one neighbour with lowest cost in boundary (R5), calculate cost for neighbour R8



Selected

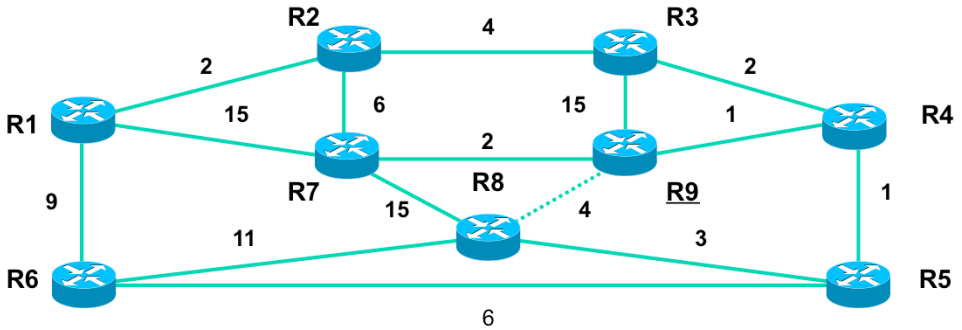
R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1
R5	9	R4

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			
R9	9	R4	R8	12	R5						

 IP Routing, v6.2 129

Select router with lowest cost in boundary (R9), calculate cost for neighbours R8



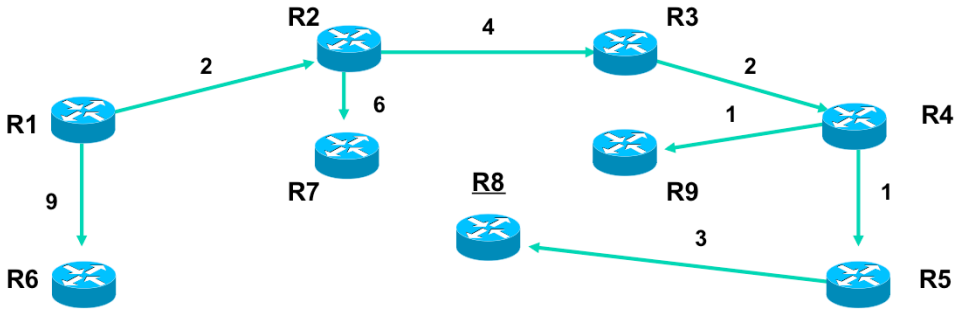
Selected

R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1
R5	9	R4
R9	9	R4

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			
R9	9	R4	R8	12	R5						
R8	12	R5									

Select last router in boundary (R8), algorithm terminated, all shortest paths found



Selected

R1	0	R1
R2	2	R1
R3	6	R2
R7	8	R2
R4	8	R3
R6	9	R1
R5	9	R4
R9	9	R4
R8	12	R5

Boundary

R2	2	R1	R6	9	R1	R7	15	R1			
R6	9	R1	R7	8	R2	R3	6	R2			
R6	9	R1	R7	8	R2	R9	21	R3	R4	8	R3
R6	9	R1	R4	8	R3	R9	10	R7	R8	23	R7
R6	9	R1	R8	23	R7	R9	9	R4	R5	9	R4
R9	9	R4	R8	20	R6	R5	9	R4			
R9	9	R4	R8	12	R5						
R8	12	R5									

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Creating the Database

- **The basic means for creating and maintaining the database are the so-called**

  Link States

- **A link state stands for an intact (synchronized) local neighbourhood between two routers**
  - The link state is created by these two routers
  - Other routers are notified about this link state via a special broadcast-mechanism ("traffic-news")
    - Flooding together with sequence numbers stored in topology database
  - Link states are verified continuously

How are Link States used?

- **Adjacent routers declare themselves as neighbours by setting the link state up (or down otherwise)**
  - The link-state can be checked with hello messages
    - Note: Link state down is no explicitly expressed, it is just the absence of the link to the former neighbour in the LSA announcement
- **Every link state change is published to all routers of the OSPF domain using Link State Advertisements (LSAs)**
  - Is a broadcast mechanism
  - LSAs are much shorter than routing tables
    - Because LSAs contain only the actual changes
    - That's why distance vector protocols are much slower
  - Whole topology map relies on correct generation and delivery of LSAs
    - Synchronization of a distributed database !!!

OSPF Communication Principle 1

- **OSPF messages are transported by IP**
  - ip protocol number 89
- **During initialization a router sends hello-messages to all directly reachable routers**
  - To determine its neighbourhood
  - Can be done automatically in broadcast networks and point-to-point connections by using the IP multicast-address 224.0.0.5 (all OSPF routers)
  - Non-broadcast networks: configuration of the neighbourhood-routers is required (e.g. X25)
- **This router also receives hello-messages from other routers**

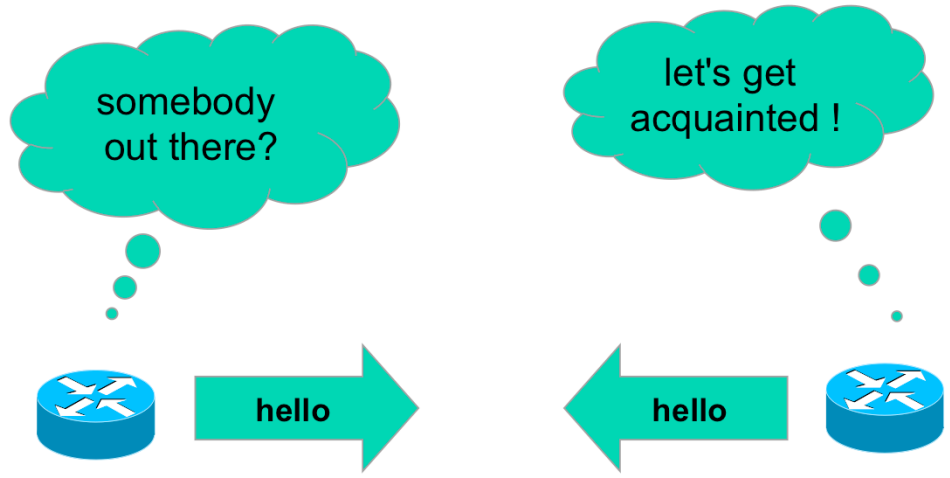
OSPF Communication Principle 2

- **Each two acquainted routers send database description messages to each other, in order to publish their topology database**
- **Unknown or old entries are updated via link state request and link state update messages**
  - Which synchronizes the topology databases
- **After successful synchronization both routers declare their neighbourhood (adjacency) via router LSAs (using link state update messages)**
  - Distributed across the whole network

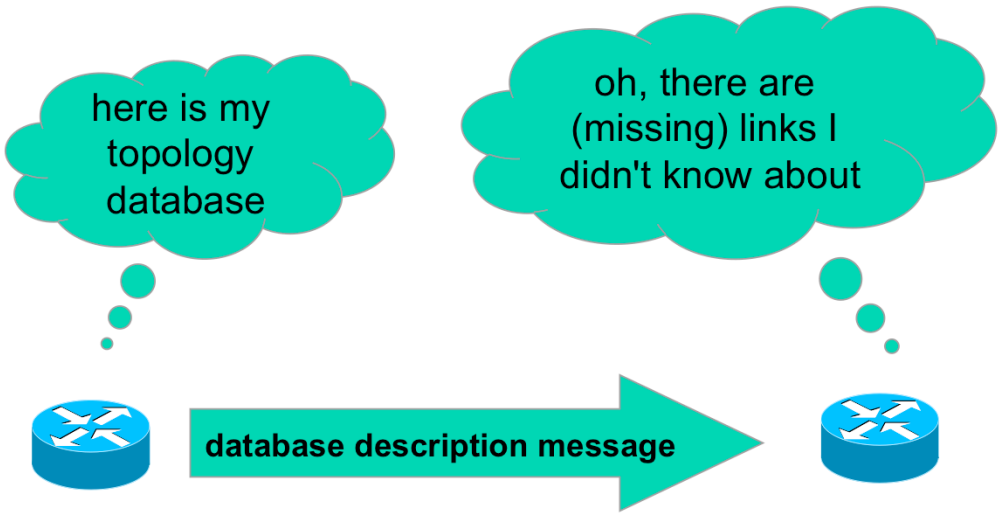
OSPF Communication Principle 3

- **Periodically, every router verifies its link state to its adjacent neighbours using hello messages**
- **From now only changes of link states are distributed**
  - Using link state update messages (LSA broadcast-mechanism)
- **If neighbourhood situation remains unchanged, the periodic hello messages represents the only routing overhead**
  - Note: additionally all Link States are refreshed every 30 minutes with LSA broadcast mechanism

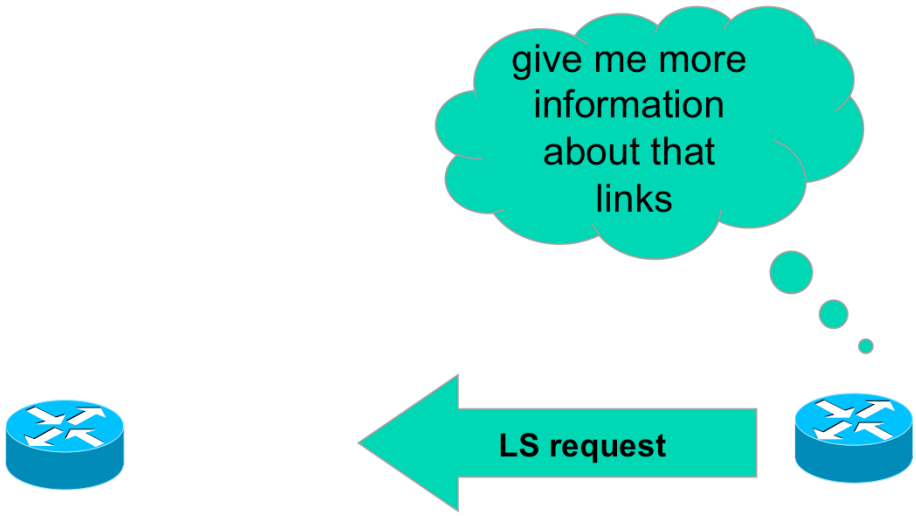
OSPF Communications Summary 1



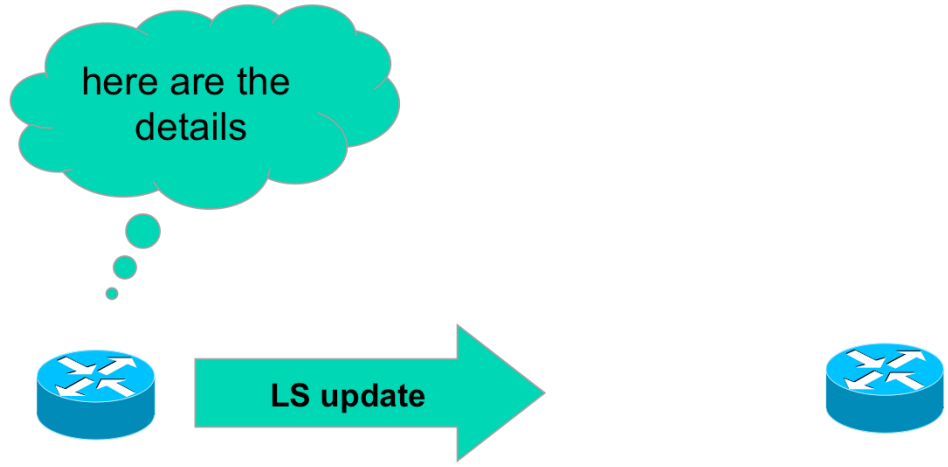
OSPF Communications Summary 2



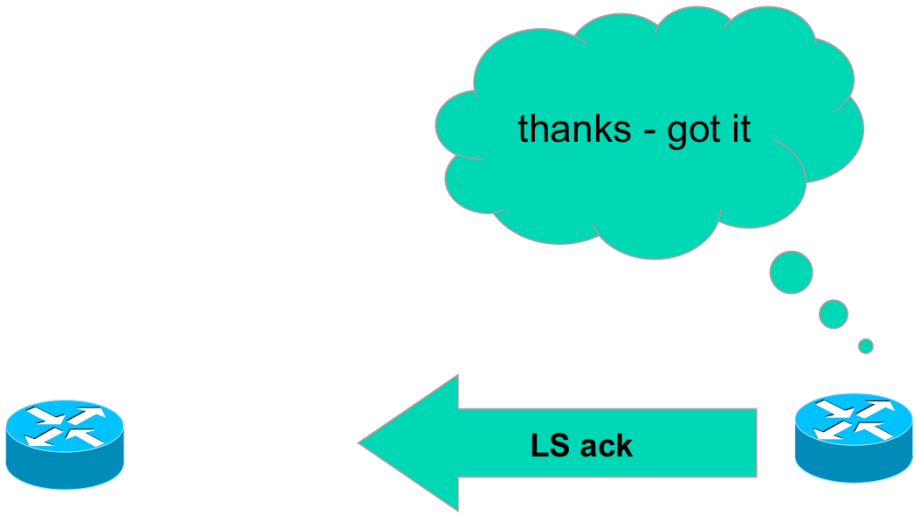
OSPF Communications Summary 3



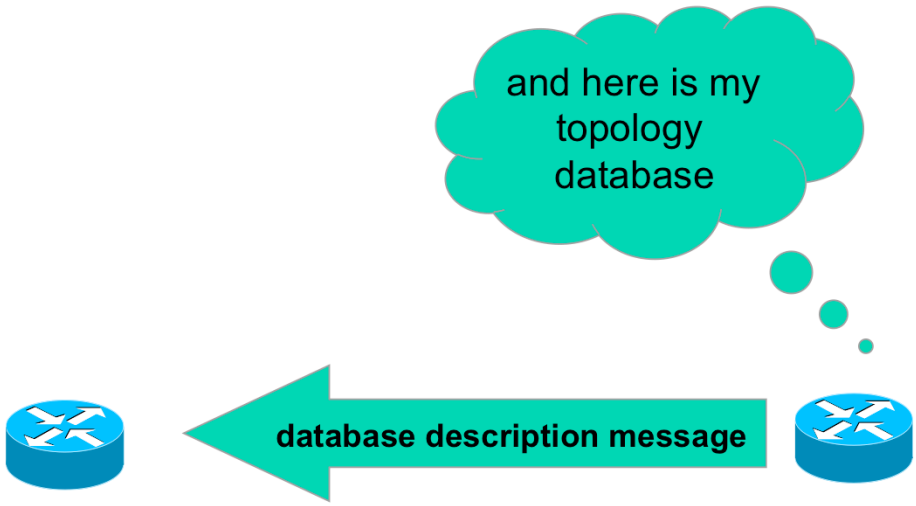
OSPF Communications Summary 4



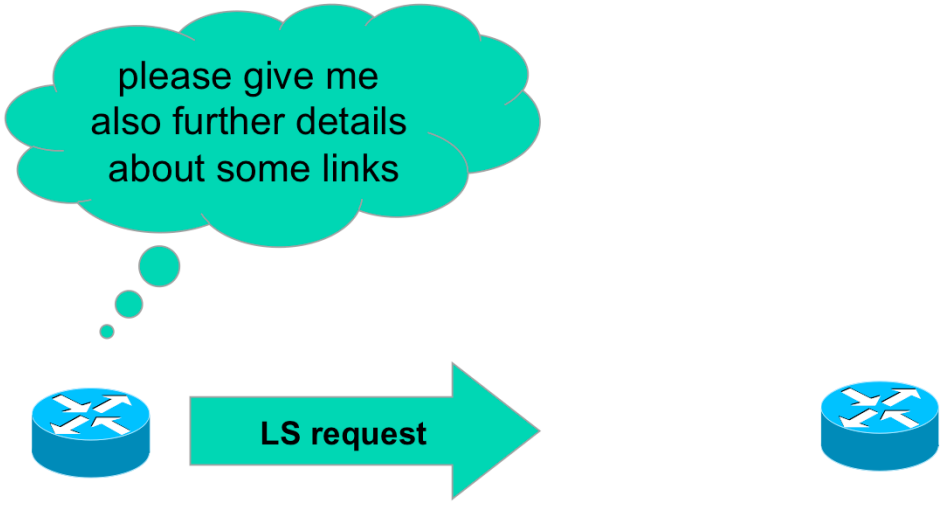
OSPF Communications Summary 5



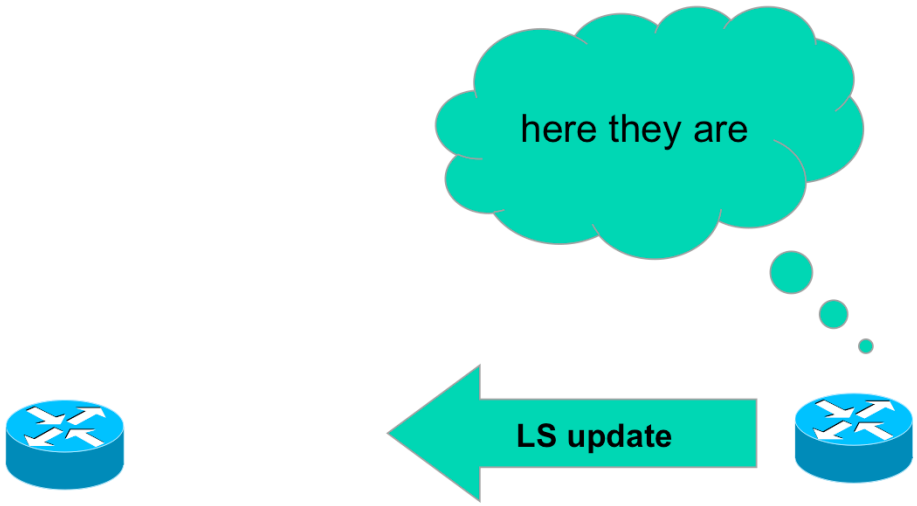
OSPF Communications Summary 6



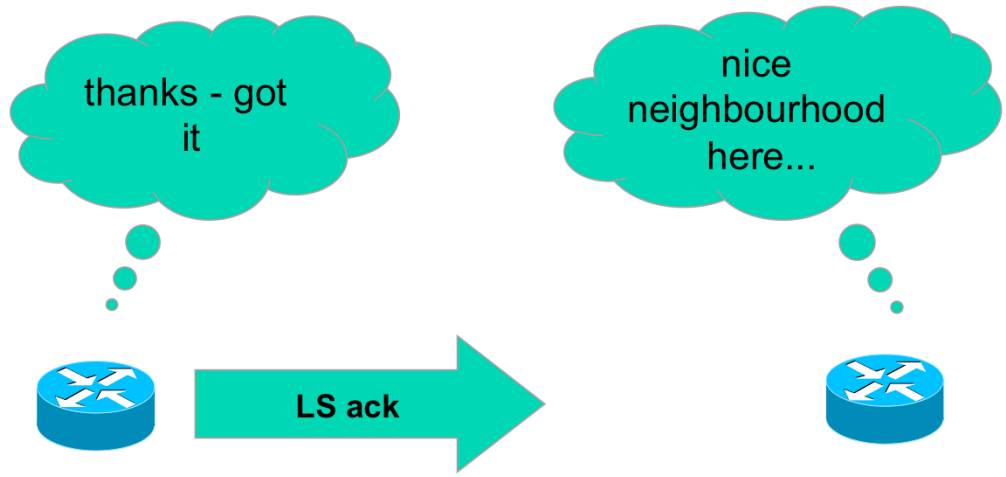
OSPF Communications Summary 7

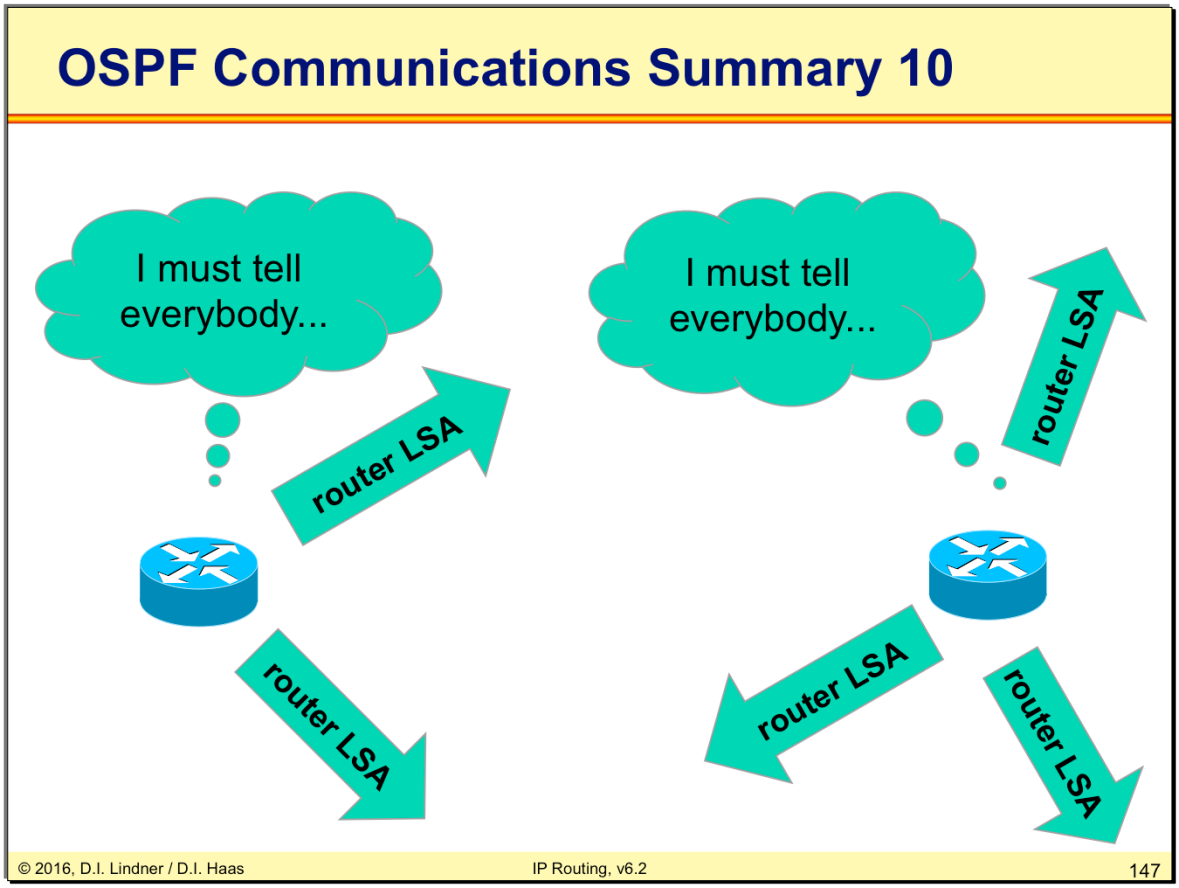


OSPF Communications Summary 8

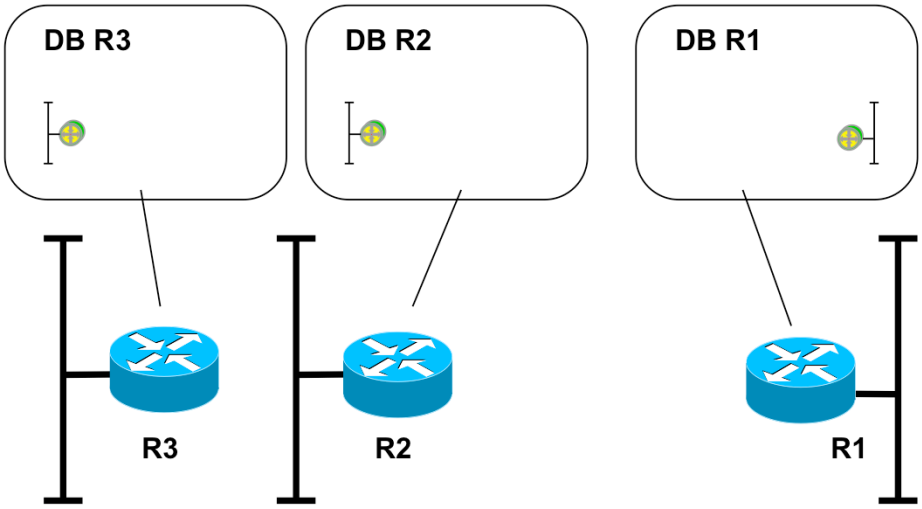


OSPF Communications Summary 9



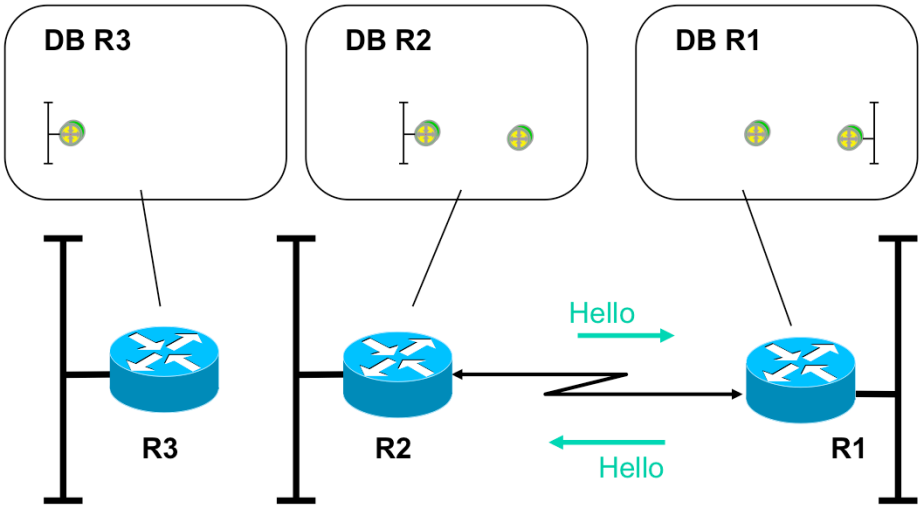

OSPF Communications Summary 10
I must tell everybody...
router LSA
router LSA
I must tell everybody...
router LSA
router LSA
router LSA
© 2016, D.I. Lindner / D.I. Haas
IP Routing, v6.2
147

OSPF Start-up



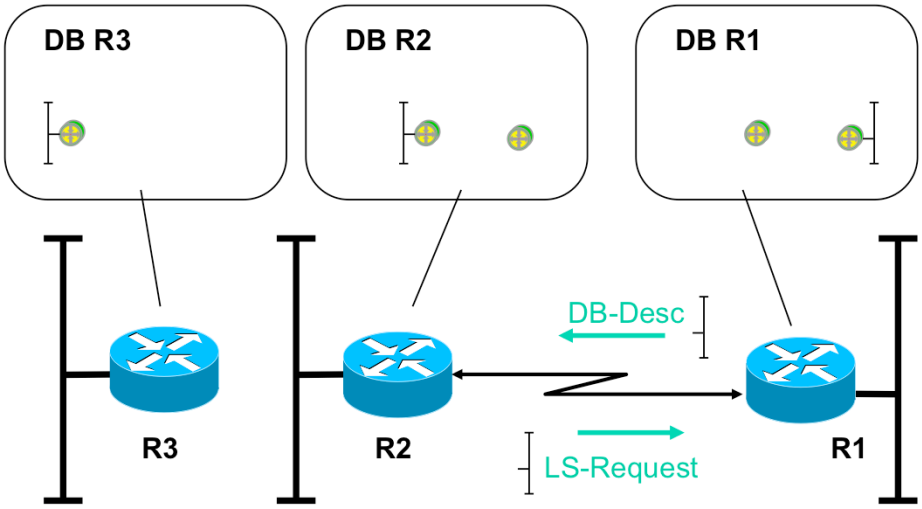
starting position: all routers initialized,
no connection between R1-R2 or R2-R3

OSPF Hello R1 - R2



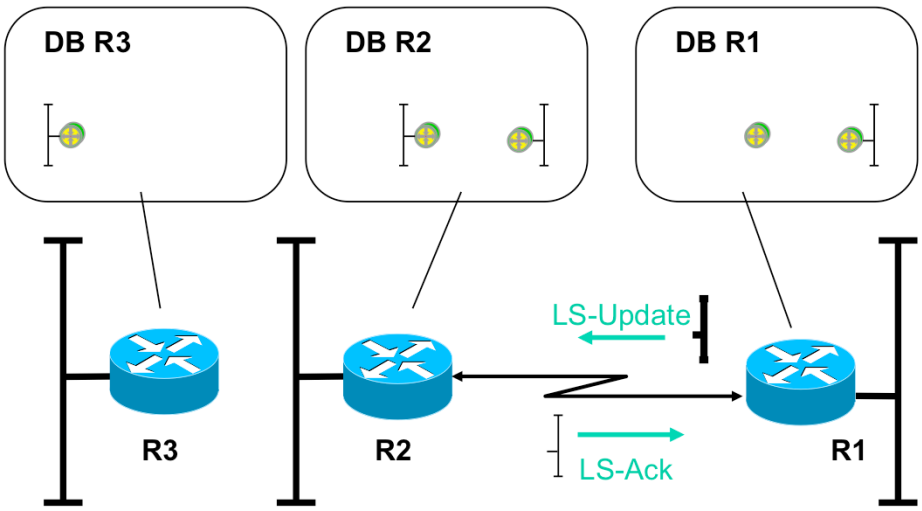
link between R1-R2 activated: get acquainted using hello messages

OSPF Data Base Description R1 -> R2



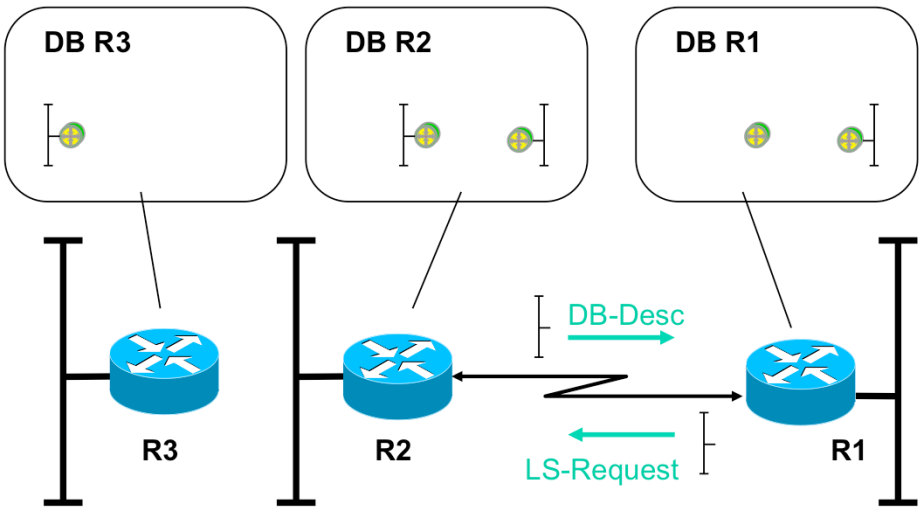
database synchronization: R1 master sends Database-Description, R2 slave sends Link-State Request

OSPF Data Base Update R1 -> R2



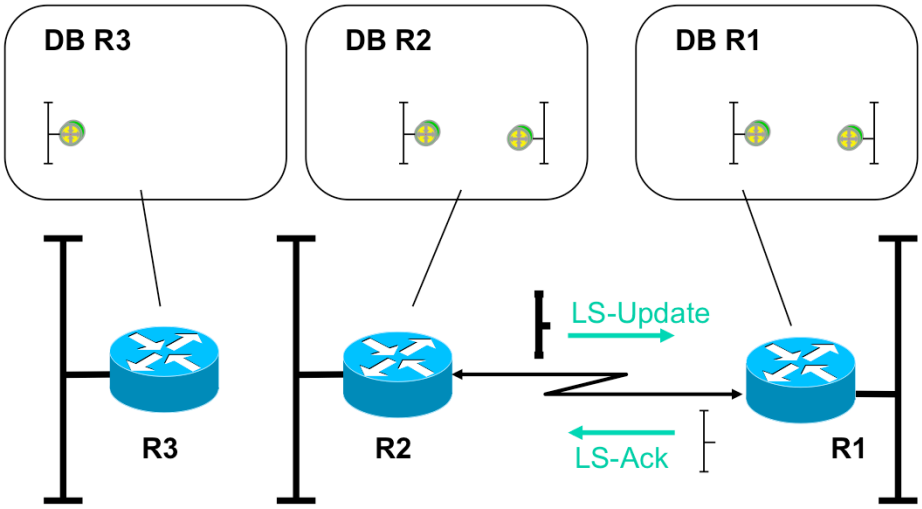
database synchronization: R1 master sends Link-State Update, R2 slave sends Link-State Acknowledgement

OSPF Data Base Description R2 -> R1



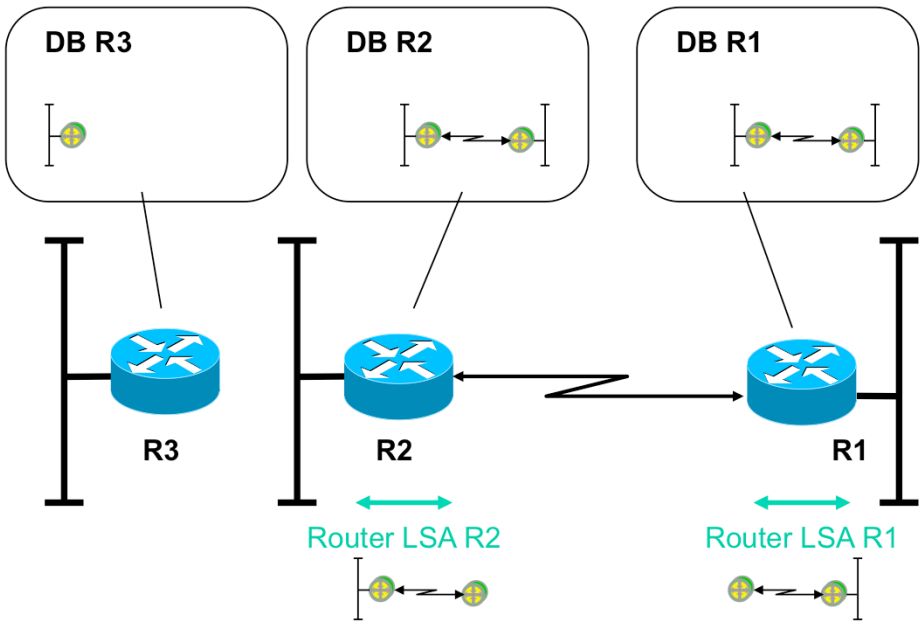
database synchronization: R2 master sends Database-Description, R1 slave sends Link-State Request

OSPF Data Base Update R2 -> R1



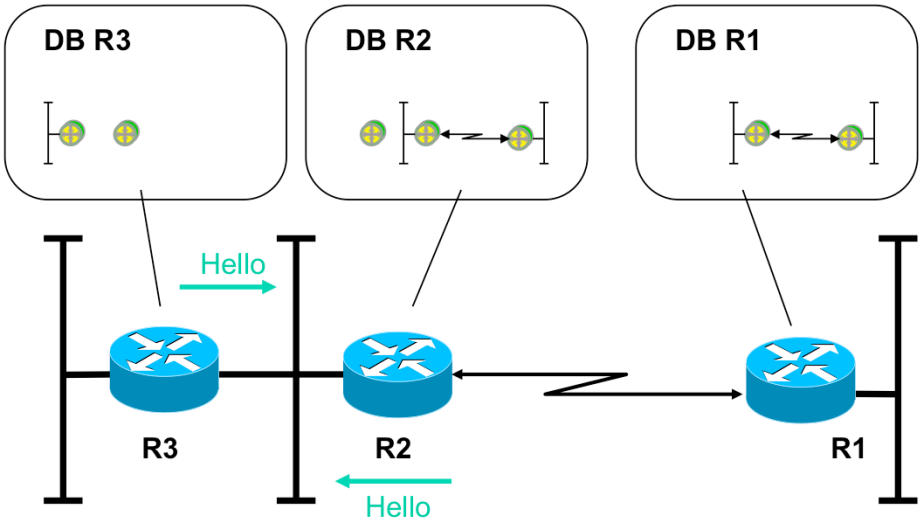
database synchronization: R2 master sends Link-State Update, R1 slave sends Link-State Acknowledgement

OSPF Router LSA Emission



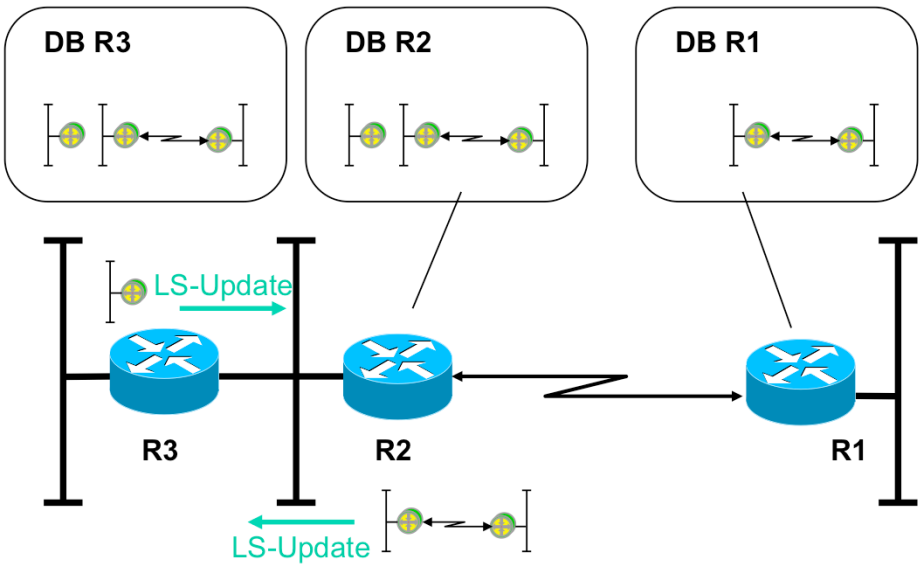
R1 and R2 have synchronized their database completely and notify other nodes about their links

OSPF Hello R2 - R3



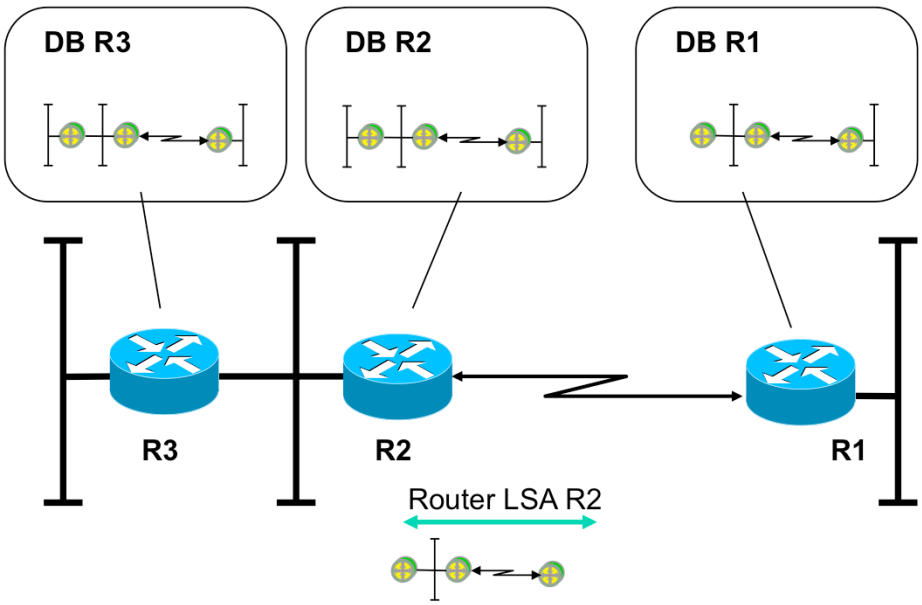
link between R2-R3 activated: get acquainted using Hello, determination of designated router

OSPF Database Update



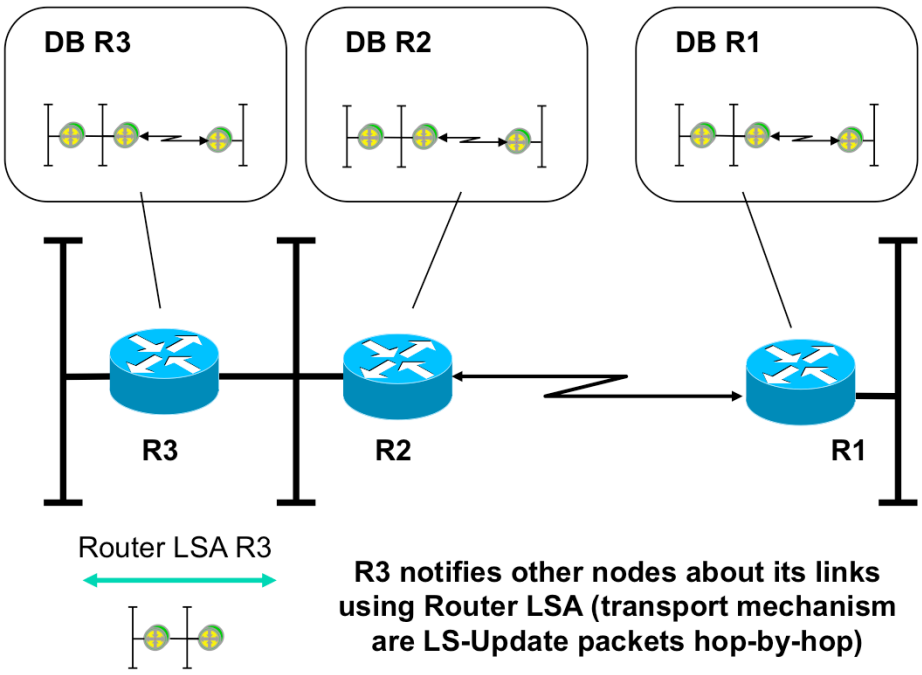
R2 and R3 synchronize their databases (DB-Des., LS-Req.,LS-Upd., LS-Ack.)

OSPF Router LSA Emission R2



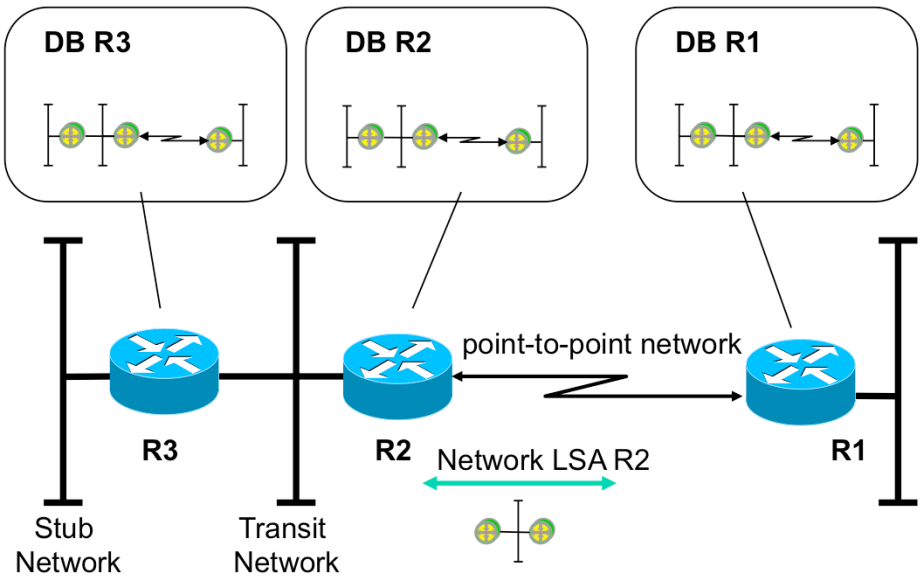
R2 notifies other nodes about its links using Router LSA,
(transport mechanism are LS-Update packets hop-by-hop)

OSPF Router LSA Emission R3



R3 notifies other nodes about its links using Router LSA (transport mechanism are LS-Update packets hop-by-hop)

OSPF Network LSA R2



Designated Router R2 notifies other nodes about the multi-access network using Network-LSA (transport mechanism are LS-Update packets hop-by-hop

Agenda

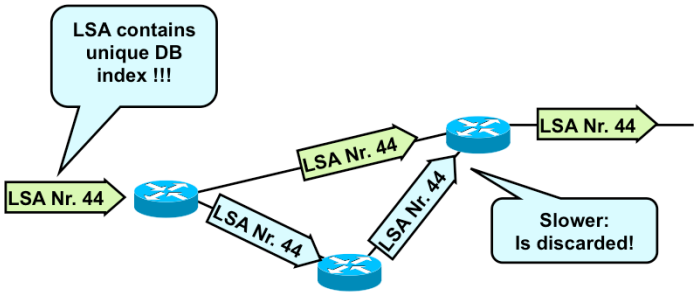
- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

LSA Broadcast Mechanism (1)

- **Flooding mechanism**
  - Receive of LSA on incoming interface
  - Forwarding of LSA on all other interfaces except incoming interface
  - Well known principle to reach all parts of a meshed network
    - Remember: Transparent bridging – Ethernet switching for unknown destination MAC address
  - "Hot-Potato" method
- **Avoidance of broadcast storm:**
  - With the help of LSA sequence numbers carried in LSA packets and topology database
    - Remember: In case of Ethernet switching we had STP to avoid the broadcast storm
    - În our case we want to establish topology database so we do not have any STP information; SPF information and hence routing tables will result from existence of consistent topology databases
    - "Chicken-Egg" problem

LSA Sequence Number

- **In order to stop flooding, each LSA carries a sequence number**
- **Only increased if LSA has changed**
  - So each router can check if a particular LSA had already been forwarded
  - To avoid LSA storms
- **32 bit number**



 IP Routing, v6.2 162

When reaching the end of the 32 bit sequence number the associated router will wait for an hour so that this LSA ages out in each link state database. Then the router resets the sequence number (lowest negative number i. e. MSB=1, 80000001) and continues to flood this LSA.

Each LSA carries also a 16 bit age value, which is set to zero when originated and increased by every router during flooding. LSAs are also aged as they are held in each router's database. If sequence numbers are the same, the router compares the ages the younger the better but only if the age difference between the recently received LSA is greater than MaxAgeDiff; otherwise both LSAs are considered to be identical.

Note:

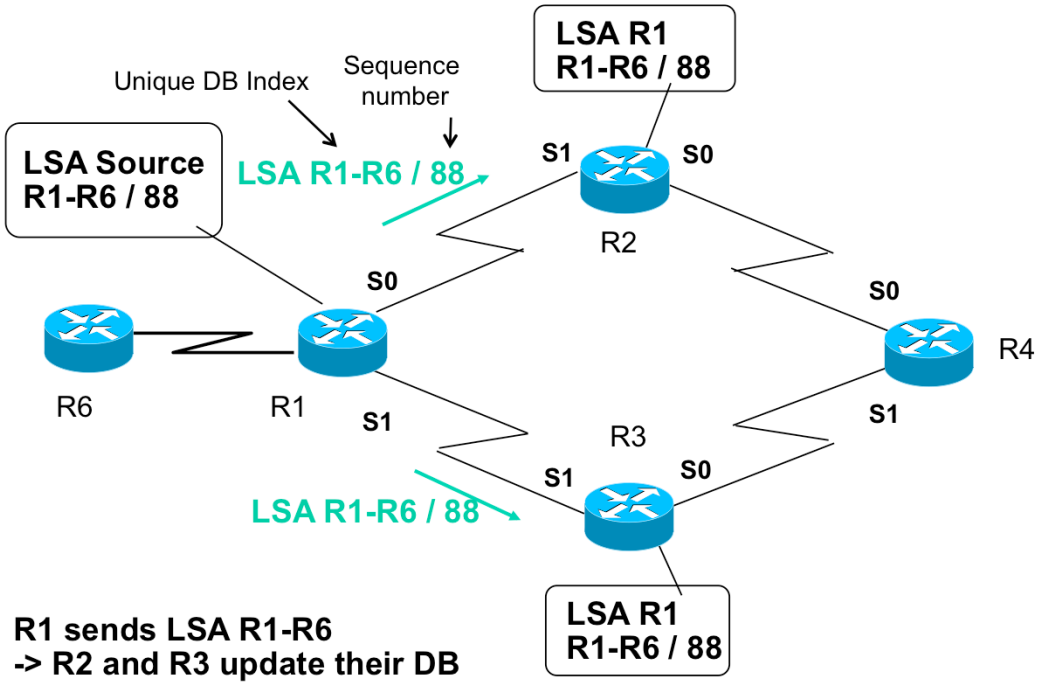
Radia Perlman proposed a "Lollipop" sequence number space but today a linear space is used as described above.

Since signed integers are used to describe sequence numbers, 8000001 represents the most-negative number in a hexadecimal format. To verify this, the 2-complement of this number must be calculated. This can be done in two steps. First calculate the 1-complement by simply inverting the binary number, that is the most significant byte "0x80" which is "1000 000" is transformed to "0111 111", the least significant byte "0x01" which is "0000 0001" is transformed to "1111 1110" and all other bytes in between are now "1111 1111". Secondly, in order to receive the 2-complement, "1" must be added. Then the final result is "0111 1111 1111 1111 1111 1111 1111 1111", which is the absolute number (without sign).

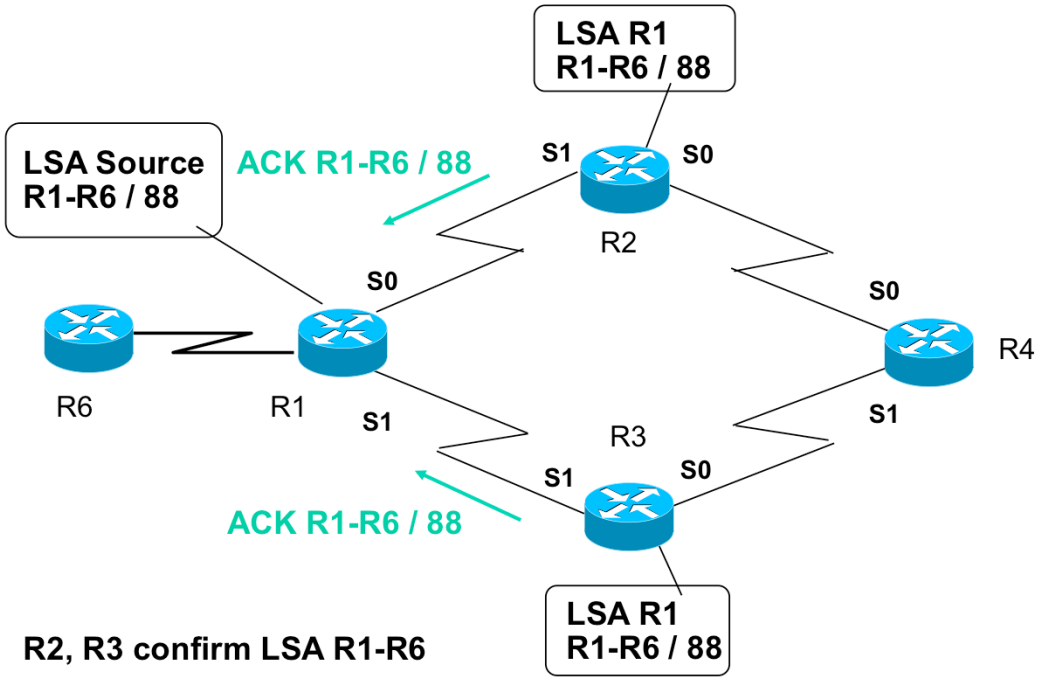
LSA Broadcast Mechanism (2)

- **LSA must be safely distributed to all routers within an area (domain)**
  - Consistency of the topology-database depends on it
  - Every LS-Update is acknowledged explicitly (using LS-ACK) by the neighbor router
  - If a LS-ACK stays out, the LS-Update is repeated (timeout)
  - If the LS-ACK fails after several trials, the adjacency-relation (the link state between the routers) is cleared

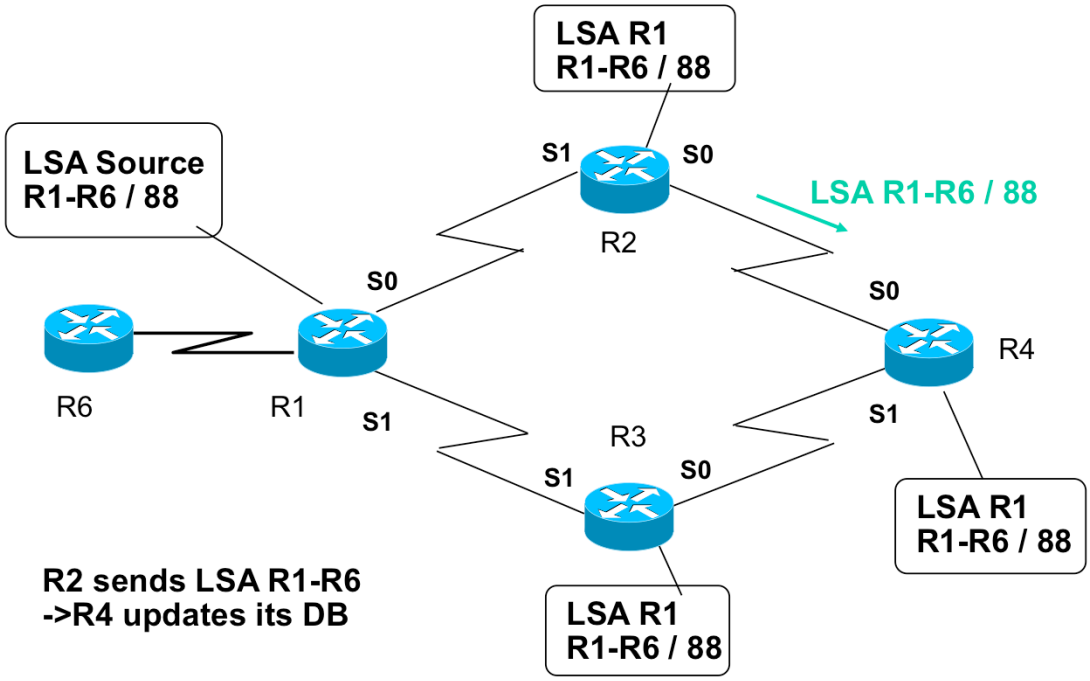
LSA Broadcast Example (1)



LSA Broadcast Example (2)

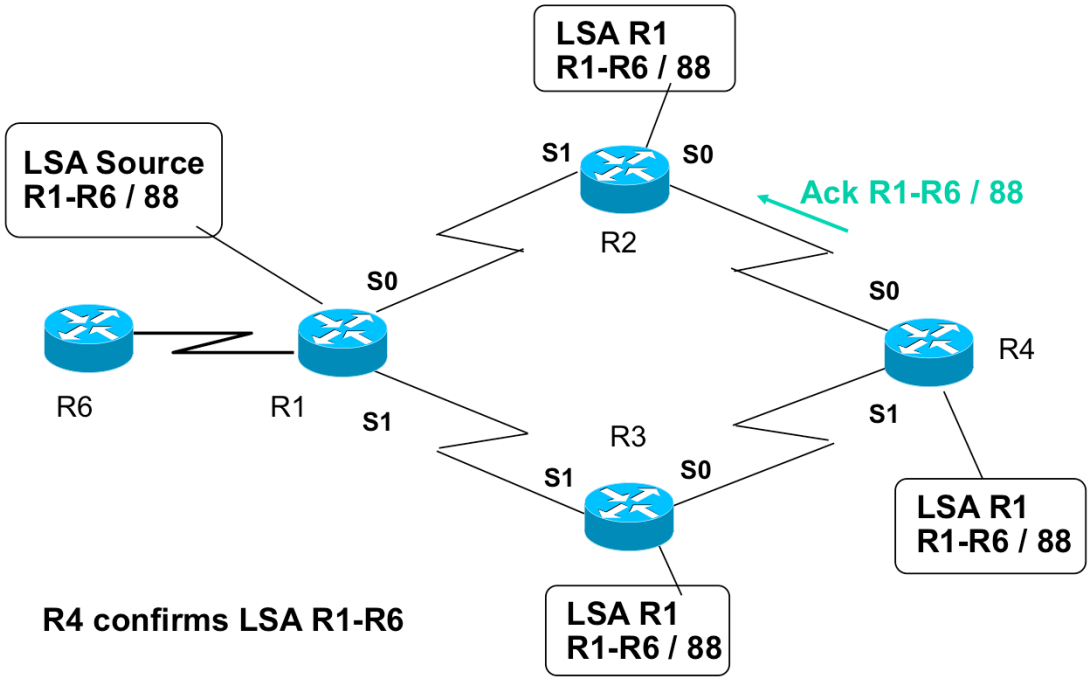


LSA Broadcast Example (3)

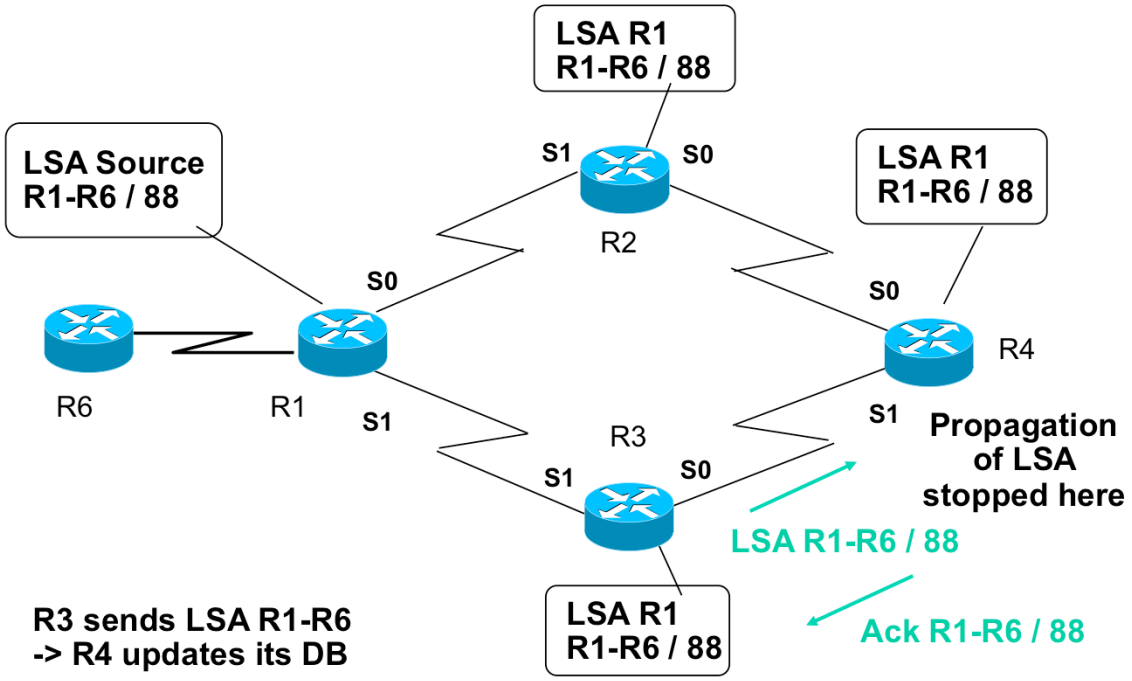


 IP Routing, v6.2 166

LSA Broadcast Example (4)



LSA Broadcast Example (5)



 IP Routing, v6.2 168

LSA Usage

- **Additionally, link states are repeated every 30 minutes to refresh the databases**
  - Link states – if not refreshed - become obsolete after 60 minutes and are removed from the databases
- **Reasons:**
  - Automatic correction of unnoticed topology-mistakes (e.g. happened during distribution or some router internal failures in the memory)
  - Combining two separated parts of an OSPF area (here OSPF also assures database consistency without intervention of an administrator)

How are LSA unique?

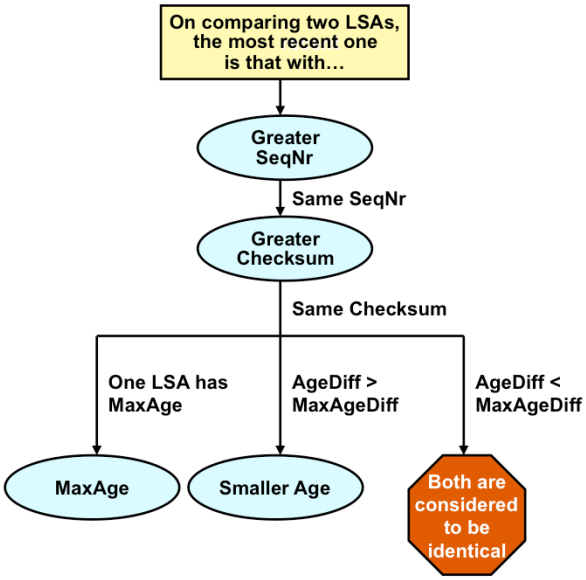
- **Each router as a node in the graph (link state topology database)**
  - Is identified by a unique Router-ID
  - Note: automatically selected on Cisco routers
    - Either numerically highest IP address of all loopback interfaces
    - Or if no loopback interfaces then highest IP address of physical interfaces
- **Every link and hence LS between two routers**
  - Can be identified by the combination of the corresponding Router-IDs
  - Note:
    - If there are several parallel physical links between two routers the Port-ID will act as tie-breaker

Note that loopback interfaces are more stable than any physical interface. Furthermore it's easier for an administrator to manage the network using loopback addresses for Router-IDs.

Detailed Flooding Decisions

FYI

- **LSA is identified by its**
  - LS type
  - Link State ID
  - Advertising Router
- **The most recent one of two instances of the same LSA is determined by:**
  - LS sequence number
  - LS checksum
  - LS age
- **MaxAgeDiff (15 min) as tolerance value**



 IP Routing, v6.2 171

Each LSA carries also a 16 bit age value, which is set to zero when originated and increased by every router during flooding. LSAs are also aged as they are held in each router's database. If sequence numbers are the same, the router compares the ages the younger the better but only if the age difference between the recently received LSA is greater than MaxAgeDiff; otherwise both LSAs are considered to be identical.

LS Age

FYI

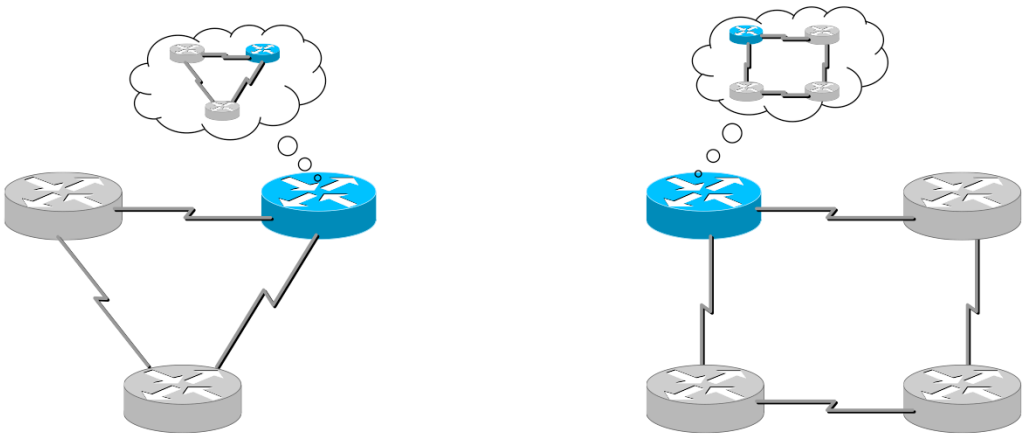
- Originating router sets LS age = 0 seconds
- Increased during flooding by InfTransDelay by every router
- Also increased while stored in database
- Age is never incremented past MaxAge (60 min)
- LSAs having MaxAge:
  - Are not used in routing table calculation anymore
  - Are reflooded immediately
  - Are always considered as most recent
  - Thus quickly flushed from routing domain
- Responsible router maintains LSRefreshTime (30 min) to refresh LSAs periodically

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Basic Principle (1)

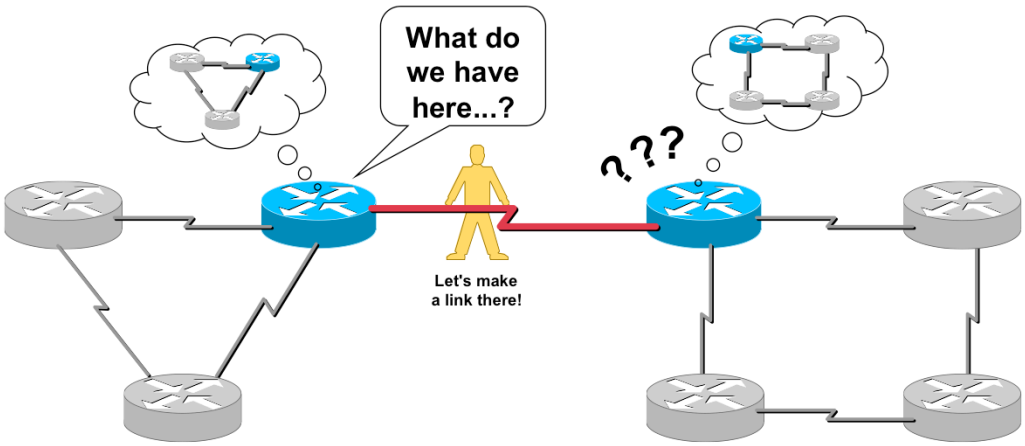
- **Consider two routers, lucky integrated in their own networks...**



The routers on the slide have 2 stable networks, there are no periodic link state updates, just hello messages.

Basic Principle (2)

- Suddenly, some brave administrator connects them via a serial cable...
- Both interfaces are still in the "Down state"

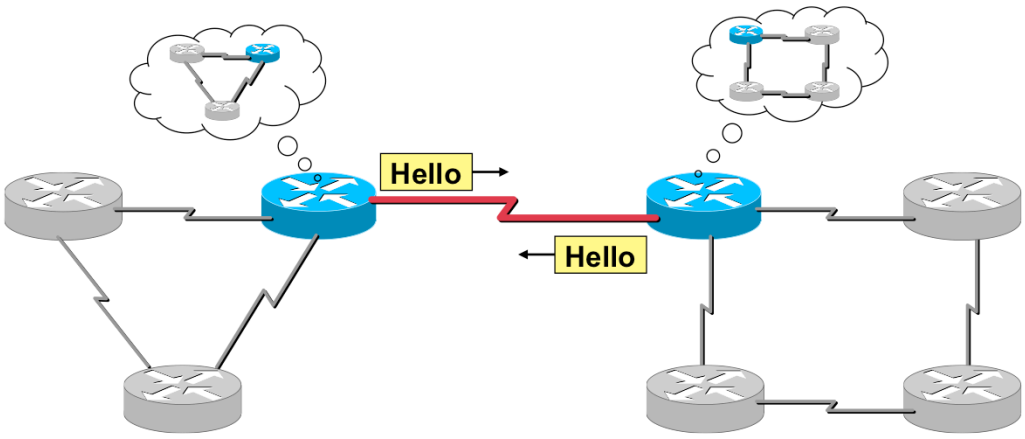


 IP Routing, v6.2 175

After the link is connected, the routers detect a new network (OSPF is configured on the interface and interfaces are enabled).

Basic Principle (3)

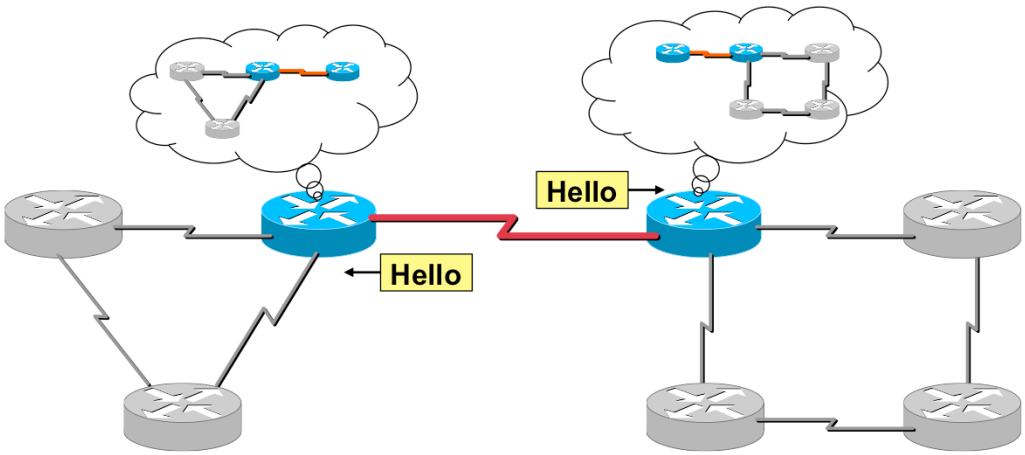
- **Init state:**
  - Friendly as routers are, they welcome each other using the "Hello protocol"…



OSPF routers send Hello packets out all OSPF enabled interfaces on a multicast address 224.0.0.5. Then the router waits for a reply (another hello from the other side) which must arrive within 4 x hello interval, otherwise the router falls back to the down state again. That is, the init state lasts only up to 4 times the hello interval.

Basic Principle (4)

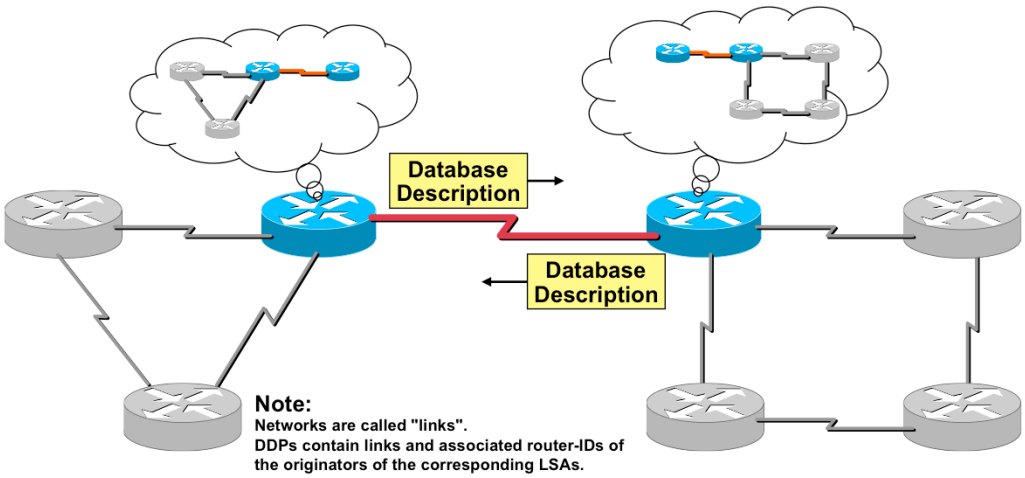
- **Two-way state:**
  - Each Hello packet contains a list of all neighbors (IDs)
  - Even the two routers themselves are now listed (=> 2-way state condition)
  - Both routers are going to establish the new link in their database...



If two routers sharing a common link and they agree on a certain parameters in their respective Hello packets, they will become neighbors.

Basic Principle (5)

- **Exstart state:**
  - Determination of master (highest IP address) and slave
  - Needed for loading state later
- **Exchange state:**
  - Both router start to offer a short version of their own roadmap, using "Database Description Packets" (DDPs)
  - DDPs contain partial LSAs, which summarize the links of every router in the neighbor's topology table.



 IP Routing, v6.2 178

After neighborship is established, the routers enter the "exstart state" and determine who of them is master and who is slave. This will be needed later as the master will begin to send LS-Request packets. The rule is simple: the router with the highest IP address (of the two involved interfaces on that link) is master.

Then, both routers enter the exchange state and exchange database description packets (DDPs), which contain partial LSAs and therefore can be regarded as a summary of their topology database.

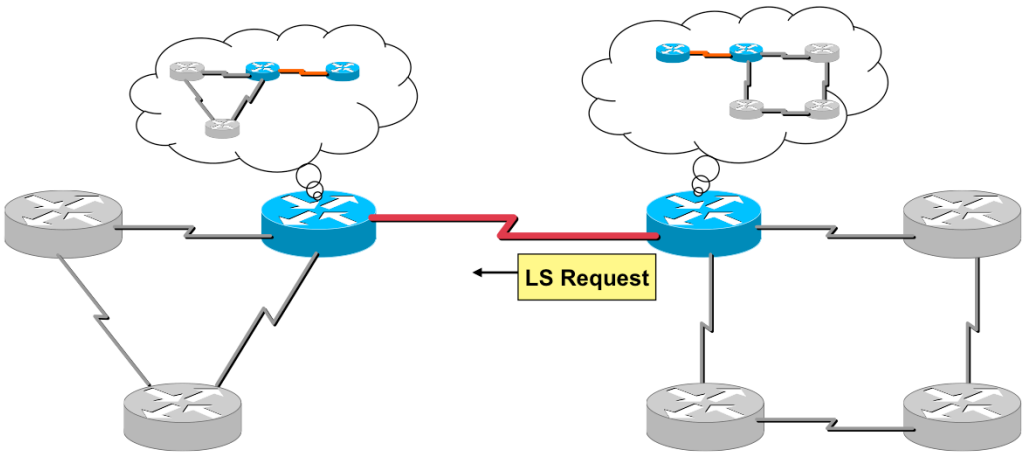
Note: typically a series of DDPs are sent from each side. Each advertised link is identified by a OSPF router ID, which represents the originator of that information.

Both routers send out a series of database description packets containing the networks held in the topology database. These networks are referred to as links. Most of the information about the links has been received from other routers (via LSAs). The router ID refers to the source of the link information.

Each link will have an interface ID for the outgoing interface, a link ID, and a metric to state the value of the path. The database description packet will not contain all the necessary information, but just a summary (enough for the receiving router to determine whether more information is required or whether it already contains that entry in its database).

Basic Principle (6)

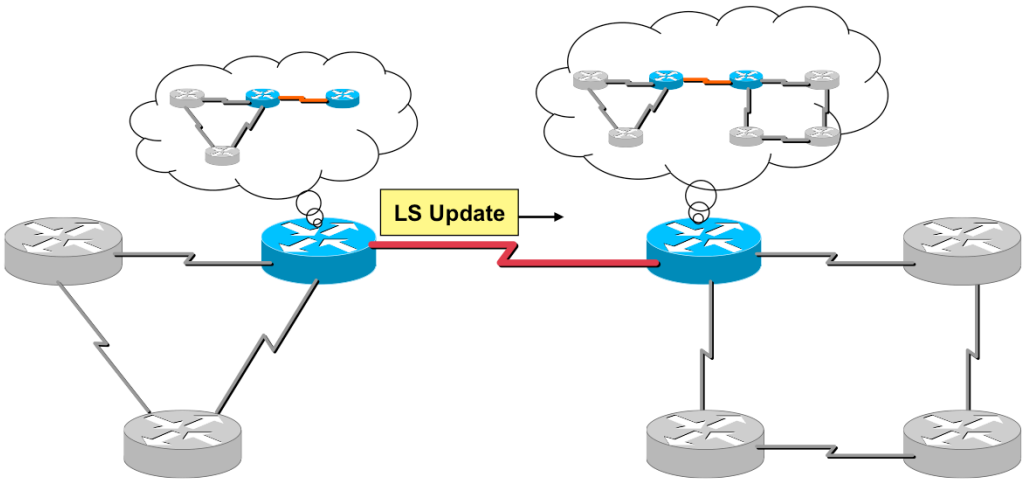
- **Loading State:**
  - One router (here the right one) recognizes some missing links and asks for detailed information using a "Link State Request" (LSR) packet...



The reciever checks its database, sees it is a new information and requests a detailed information with Link State Request packet LSR.

Basic Principle (7)

- The left router replies immediately with the requested link information, using a "Link State Update" (LSU) packet ...

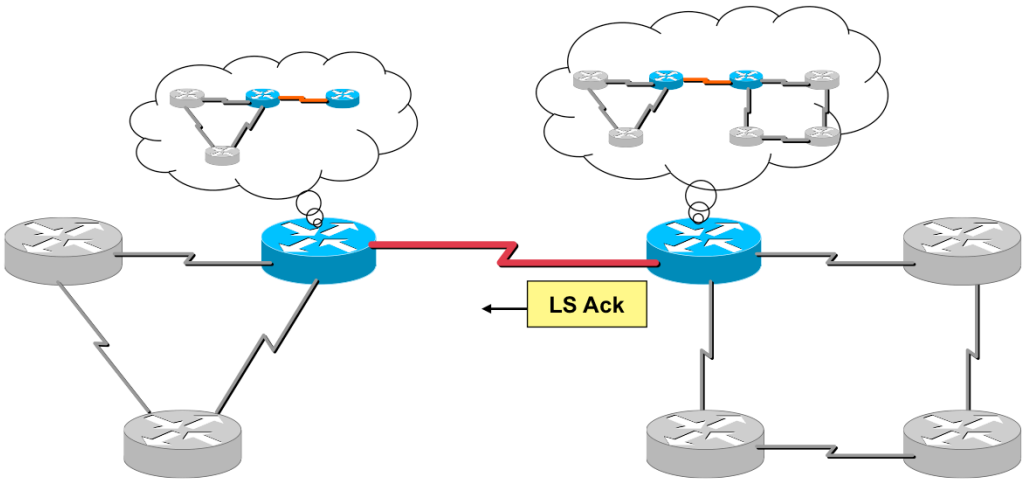


 IP Routing, v6.2 180

As a reply the left router sends a Link State Update packet LSU which contains detailed information about requested links.

Basic Principle (8)

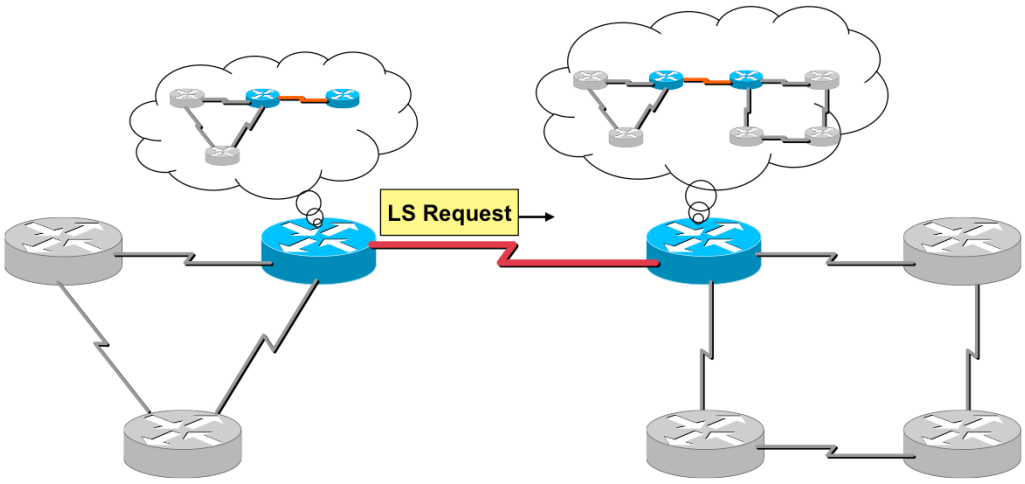
- The right router is very thankful, and returns a "Link State Acknowledgement"...



Link State Acknowledgement LSAck is used to make sure that the information is received.

Basic Principle (9)

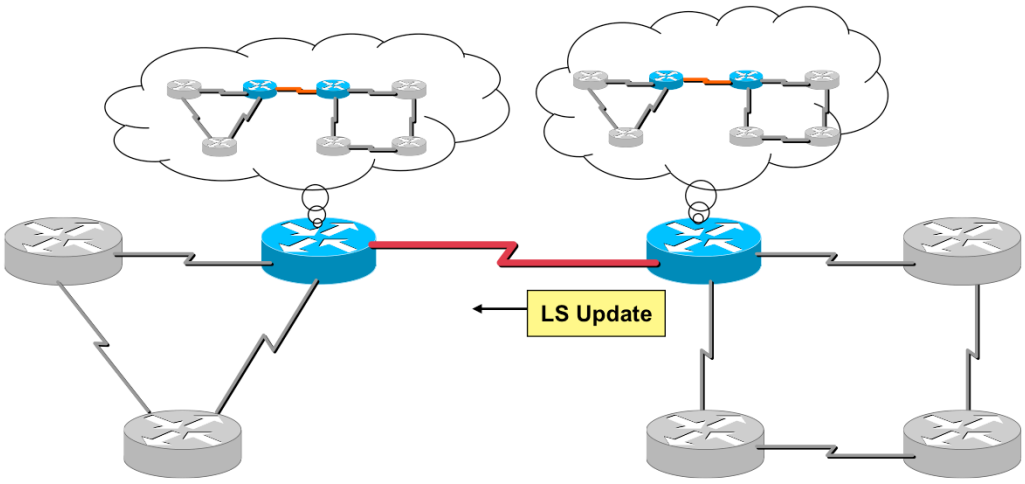
- Then the left router recognizes some unknown links and asks for further details...



LSR is sent in the other direction asking for detailed information.

Basic Principle (10)

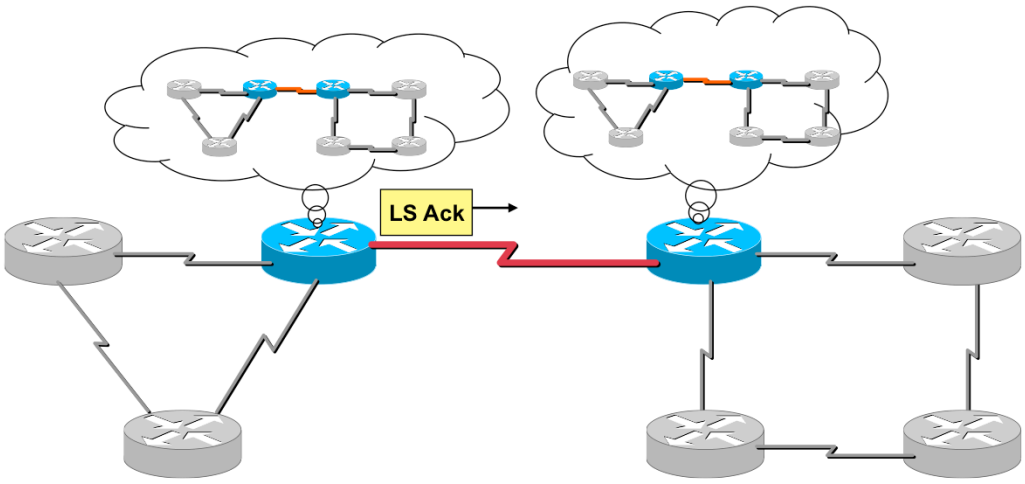
- The right router sends detailed information for the requested unknown links...



Then a LSU is sent back.

Basic Principle (11)

- **The left router replies with a link state acknowledgement – a new adjacency has been established...**
  - Neighbors are "fully adjacent" and reached the "full state"

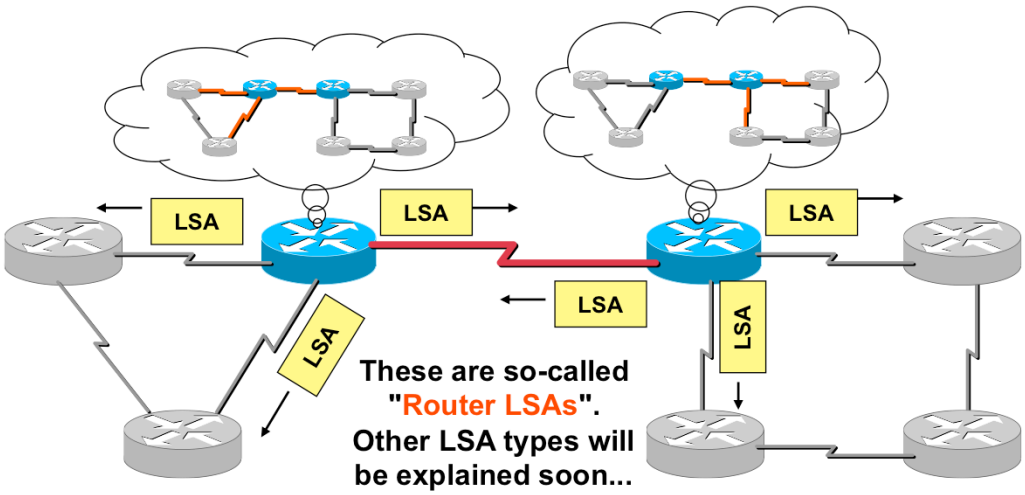


 IP Routing, v6.2 184

LSAck – saying thanks for info.

Basic Principle (12)

- Both routers tell all other routers about all local adjacencies by flooding link state advertisements (LSAs)
- Both routers now see their own IDs listed in the periodically sent Hello packets

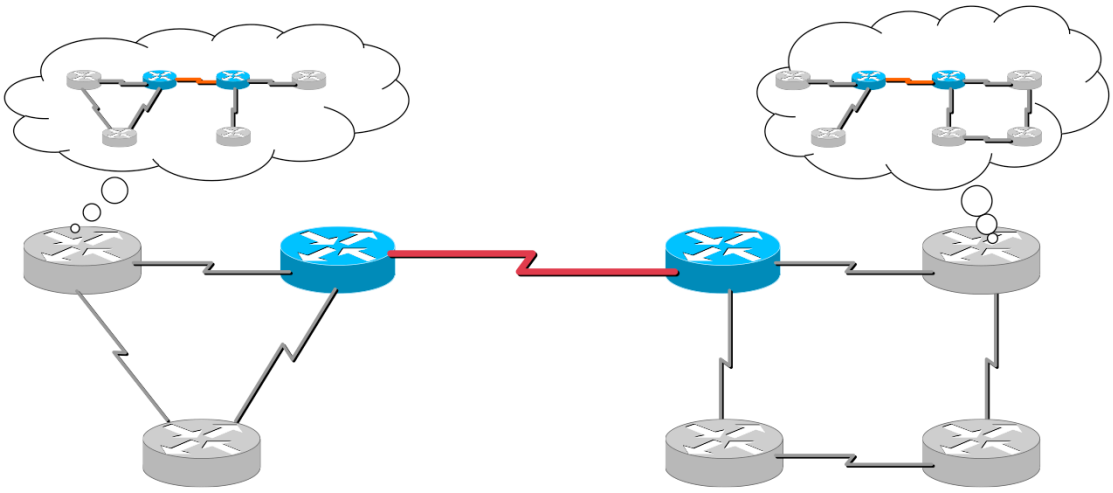


 IP Routing, v6.2 185

Now the both routers have a new information in their databases. This information is flooded to all other adjacent routers as a router LSA or LSA type 1 in which the router sends information about its own links.

Database Inconsistency

- When connecting two networks, LSA flooding only distributes information of the local links of the involved neighbors (!)



 IP Routing, v6.2 186

It might happen if you connect two existing networks together. As you can see some routers may miss a new information.

Healing Inconsistency (1)

- **Every router sends its LSAs every 30 minutes (!)**
  - Heals but long time of routing table / topology table inconsistence when combining a former split area of a OSPF domain
- **Triggering database synchronization between any two routers in the network**
  - In order to avoid long time of inconsistence
  - So whenever a router is informed by a Router-LSA about some changes in the network this router additionally will do a database synchronization with the router from which the Router-LSA was received
  - Database description packets will help to reduce traffic to the necessary minimum

Healing Inconsistency (2)

- **Optionally flash updates configured**
  - Upon receiving an LSA a router not only forwards this LSA but also immediately sends its own LSAs
  - Cisco default (can be turned off)
- **Golden OSPF design rule:**
  - Avoid splitting of an area in an OSPF environment by avoiding any single point of failures
  - Hence most parts of an area should be connected redundantly to each other

According to RFC to solve a problem each router sends a so-called refreshment LSA every 30 minutes.

Finally: Convergence!

- **When LSAs are flooded, OSPF is quiet (at least for 30 minutes)**
- **Only Hello's are sent out on every interface to check adjacencies**
  - Topology changes are quickly detected
  - Default Hello interval: 10 seconds (LAN, 60 sec WAN)
  - Hellos are terminated by neighbors

 IP Routing, v6.2 189

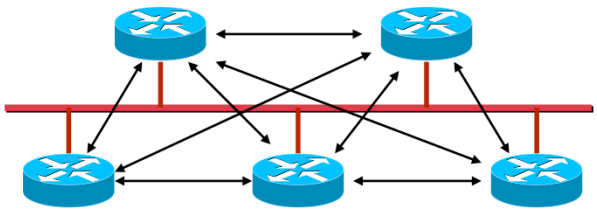
After flooding the routers are recalculating their routing tables, using SPF algorithm. There are no periodic updates like in RIP. Just Hello packets are sent every 10 seconds by default. If a router does not get a Hello from the neighbor for 40 seconds, it decides the neighbor is dead and this is a dead interval, which is 4 times the hello interval by default.

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Broadcast Multi-Access Media (1)

- When several OSPF routers have access to the same Ethernet segment they would create n(n-1)/2 adjacencies
- Furthermore, SPF algorithm requires to represent a fully meshed network as tree



Basic concept of link state requires point-to-point relationships. That fits best for point-to-point networks like serial lines but that causes a problem with shared media multi-access networks (e.g. LANs or with networks running in a so called NBMA-mode (Non Broadcast Multi Access) like X.25, Frame Relay, ATM. Hello, database description and LSA updates between each of these routers can cause huge network traffic and CPU load.

Consider the flooding process after establishment of each adjacency!!! The formation of an adjacency between every attached router would create a lot of unnecessary LSAs. A router would flood an LSA to all its adjacent neighbors, creating many copies of the same LSA on the same network.

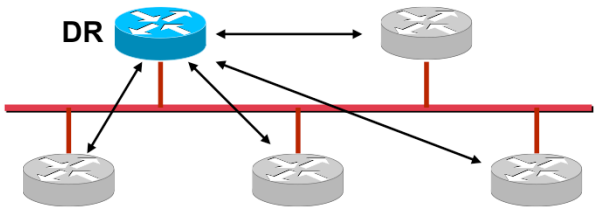
Information about all possible neighbourhood-relations seems to be redundant. The well known concept of virtual (network) node (or virtual router) is introduced to solve the problem.

Only the virtual node needs to maintain N-1 point-to-point relationship to the other nodes and hence any-to-any is not necessary.

In OSPF the virtual node is called Designated Router (DR).

Broadcast Multi-Access Media (2)

- **Solution: Elect one "Designated Router" (DR) to represent the whole LAN segment**
  - Election uses the Hello protocol
- **DR sends Network LSA**
  - List of all local routers
  - Ensures that every router on the link has the same topology database
  - Also contains subnet mask (!)
- **Each other router establishes an adjacency only to the DR**
  - Using "All DR" multicast address 224.0.0.6



 IP Routing, v6.2 192

To prevent the problems described in the previous slide, a Designated Router (DR) is elected on a multi-access network. DR is responsible for representation of the multi-access network and all the routers on it to the rest of network and management of flooding process on a multi-access network. The network itself becomes a "pseudonode" on the graph. The pseudonode is represented by the DR.

All other routers peer with the DR, which informs them of any changes on the segment.

Note: For LAN segments, the Router LSA does NOT contain the subnet mask. The subnet mask for this LAN segment is also carried inside the Network LSA.

In case of a failure the Designated Router would be single point of failure.

Therefore a Backup Designated Router (BR) is elected, too.

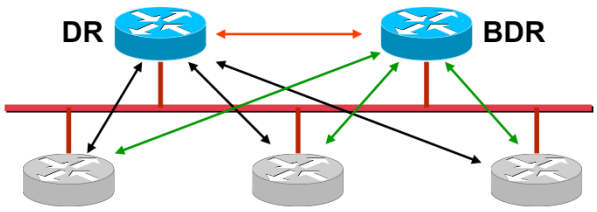
DR and BR are elected by exchanging hello-messages at start-up.

Attention !!!

The concept of DR/BDR influences only how routing information is exchanged among those routers. There is no influence on actual IP forwarding which is based on routing tables.

Broadcast Multi-Access Media (3)

- **Only the DR will send LSAs to the rest of the network**
- **For backup purposes also a Backup DR is elected (BDR)**
  - All routers also establish adjacencies to the BDR
  - BDR itself also establishes adjacency to DR



Each multi-access interface has a "Router Priority" ranging from 0 to 255 (default 1). Routers with a priority of 0 cannot become DR or BDR. The election process is performed with Hello packets which carry the priority. If some routers have the same priority, the one with the highest numerical Router ID wins. If a DR fails the BDR becomes active immediately (Hello stays out) and a new election for the BDR is started.

Note: After election of DR and BDR, adding a new router with higher priority will not replace them. The first two routers immediately become DR and BDR. The only way to control the election is to set the priority for all other routers ("DROTHER") to zero, so they cannot become DR or BDR.

DR/BDR Election Process

- **Election process starts if no DR/BDR listed in the hello packets during the init state (i. e. when two routers begin to establish an adjacency)**
  - Note: if already one DR/BDR chosen, any new router in the LAN would not change anything!
  - Therefore, the power-on order of routers is critical !!!
- **Always configure loopback interface in order to "name" your routers**
  - Loopback interface never goes down
  - Ensures stability
  - Simple to manage

It is recommended in OSPF to use the loopback interfaces for router ID. You should configure a loopback interface first and then start the OSPF process, otherwise the highest ip address from a physical interface will be taken.

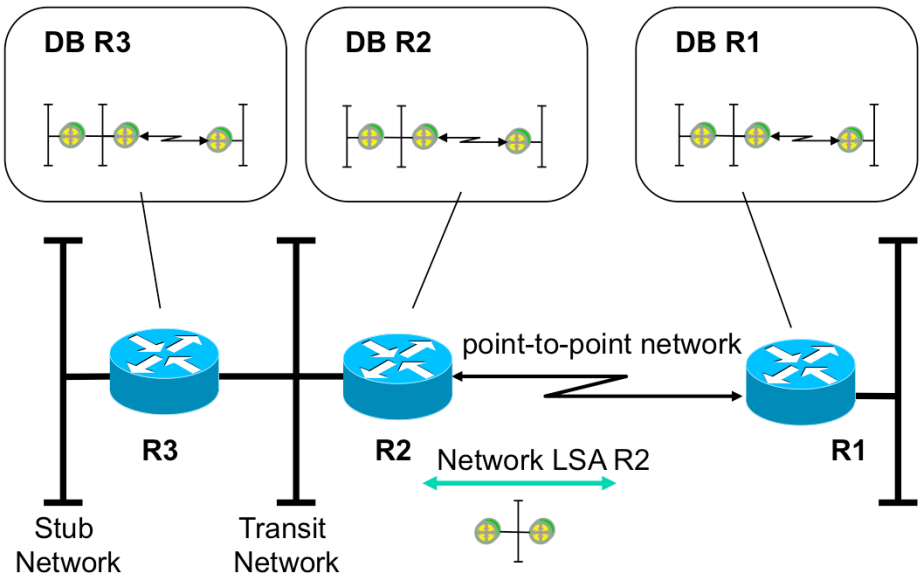
Designated and Backup Designated Router are determined using the router-priority field of the Hello message. On a DR failure, the Backup Router (BDR) continues the service.

BDR listens to the traffic on the virtual point-to-point links between all routers and the DR. Multicast addresses are used for ease that newtork sniffing. BDR recognizes a DR failure through missing acknowledge messages. Remember: Every LS-Update message requires a LS-Acknowledgement message.

DR, Router LSA, Network LSA

- **Designated Router (DR) is responsible**
  - For maintaining neighbourhood relationship via virtual point-to-point links using the already known mechanism
    - DB-Description, LS-Request LS-Update, LS-Acknowledgement, Hello, etc.
- **Router-LSA implicitly describes**
  - These virtual point-to-point links by specifying such a network as transit-network
    - Remark: Stub-network is a LAN network where no OSPF router is behind
- **To inform all other routers of domain about such a special topology situation**
  - DR is additionally responsible for emitting Network LSAs
- **Network LSA describes**
  - Which routers are members of the corresponding broadcast network

OSPF Network LSA R2



Designated Router R2 notifies other nodes about the multi-access network using Network-LSA (transport mechanism are LS-Update packets hop-by-hop

Details: OSPF Multicast Usage

- **OSPF uses dedicated IP multicast addresses for exchanging routing messages**
  - 224.0.0.5 (“All OSPF Routers")
  - 224.0.0.6 (“All Designated Routers")
- **224.0.0.5 is used as destination address**
  - By all routers for Hello-messages
    - DR and BR determination at start-up
    - link state supervision
  - By DR router for messages towards all non-DR routers
    - LS-Update, LS-Acknowledgement
- **224.0.0.6 is used as destination address**
  - By all non-DR routers for messages towards the DR
    - LS-Update, LS-Request, LS-Acknowledgement and database description messages

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

OSPF Domain / OSPF Area

- **OSPF domain can be divided in multiple OSPF areas**
  - To improve performance
  - To decouple network parts from each other
- **Performance improvement**
  - By restricting Router-LSA and Network-LSA to the originating area
    - Note: receiving a Router-LSA will cause the SPF algorithm to be performed
- **Decoupling is actually done**
  - By route summarization enabled through the usage of classless routing and careful IP address plan

 IP Routing, v6.2 199

As each link is identified by a router LSA in the OSPF database, the total OSPF routing traffic increases with the number of links and thus with the size of the network. Also the amount of network LSA will increase in larger networks. The basic idea of OSPF to overcome these limitations is to partition the whole OSPF domain into smaller "areas". The basic idea is to filter router LSAs and network LSAs on the borders between areas. Network reachabilities from outside is advertised through other LSA types. These details are discussed next.

OSPF Domain / OSPF Area

- **Every area got its own topology database**
  - Which is unknown to other areas
  - Area specific routing information stays inside this area
- **On topology changes**
  - Routing traffic causing Dijkstra's algorithm to be performed stays inside the area where the change appears
  - Route summarization reduces routing traffic drastically
- **OSPF areas are labelled with area-IDs**
  - Unique within the OSPF domain
  - Written in IP address like format or just as number
- **An OSPF domain contains**
  - At least one single area or several areas

OSPF Area Border Router

- **OSPF areas are connected by special routers**
  - Area Border Router (ABR)
- **ABR**
  - Maintains a topology database for each area he is connected to
- **All OSPF areas must be connected over a special area**
  - Backbone Area
    - Area-ID = 0.0.0.0 or area-ID = 0
  - If there is only one area in the OSPF domain this OSPF area will be the backbone area

OSPF Backbone Area

- **Non-backbone areas must not be connected directly**
  - Connection allowed only via Backbone Area
- **This OSPF rule forces**
  - A star-like topology of areas with the backbone area in the centre
- **ABRs**
  - Are connected to the backbone area by direct physical links in normal cases
  - Exception with virtual link technique if direct physical link can not be provided
    - A virtual link can be used to "tunnel" the routing traffic between an isolated area and the backbone area through another area

OSPF Routing Types 1

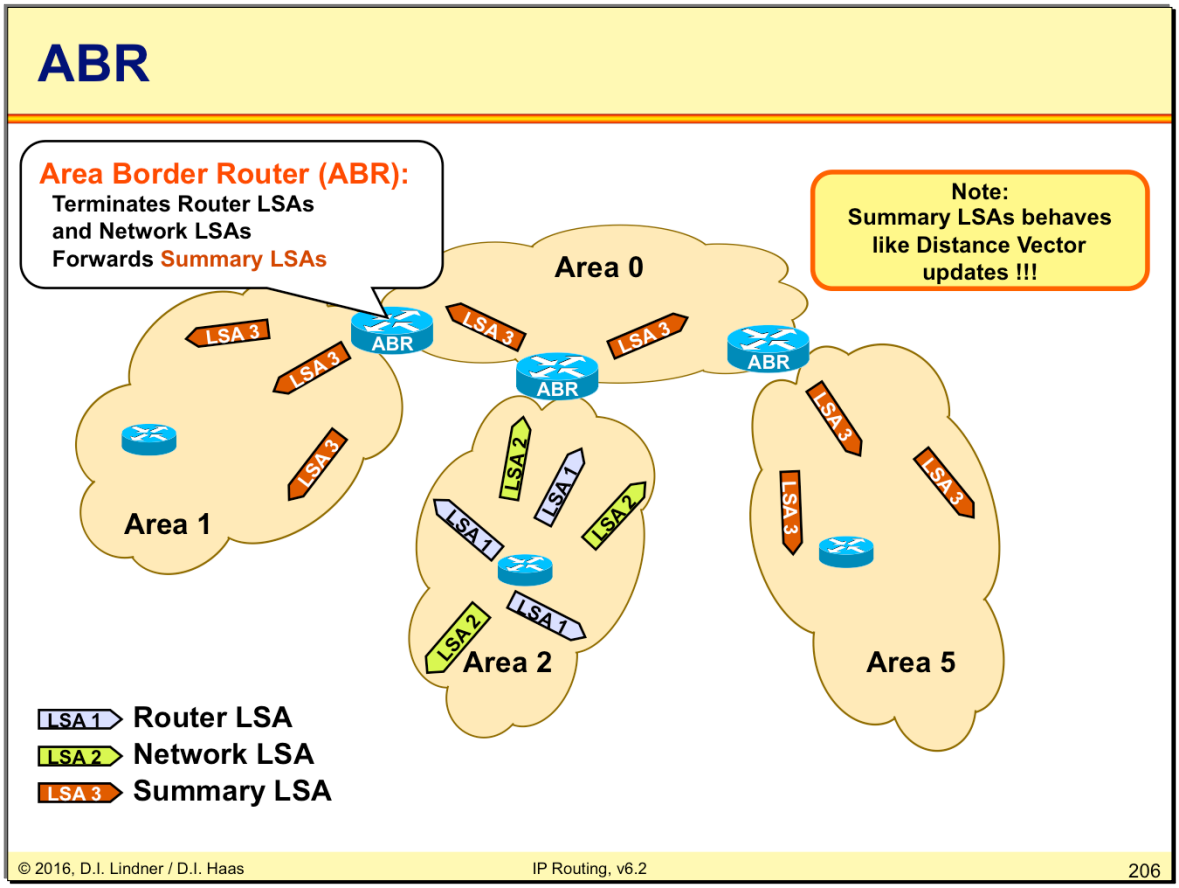
- **OSPF provides three types of routing:**
  - Intra-area routing:
    - Inside of an area (using Level 1 Router; Internal Router IR)
    - Router Link LSA (LSA type1)
    - Network Link LSA (LSA type2)
    - Note: Backbone Router is a Backbone Area Internal Router
  - Inter-area routing:
    - Between areas over a Backbone Area (using Area Border)
    - Summary Link LSA (LSA type3 and type4)
    - Type 3 to announce networks
    - Type 4 to announce IP address of ASBRs

OSPF Routing Types 2

- **OSPF provides three types of routing (cont.):**
  - Exterior routing:
    - Paths to external destinations (other AS) are configured statically or imported with EGP or BGP using Autonomous Systems Boundary Routers (ASBRs)
    - AS External Summary LSA (LSA type5) to announce external networks

Area Border Router

- **Area Border Router maintains two topology maps**
  - One for its area
  - One for the Backbone Area
- **Area Border Router exports the routes of its area to the Backbone Area**
  - Collects all topology information of its area and sends Summary LSAs to the Backbone Area
- **Area Border Router imports all routes of other areas (received from the backbone area) in its own area**
  - This is done again using Summary LSAs



Traffic from one area to another area flows through dedicated routers only, so called Area Border Routers (ABRs). The ABRs filter Router LSAs and Network LSAs. Network destinations in other areas are advertised by so-called "Network Summary LSAs", which carry simple distance-vector information i. e. which networks can be reached by which ABR.

Actually, we will deal with the following OSPF router types:

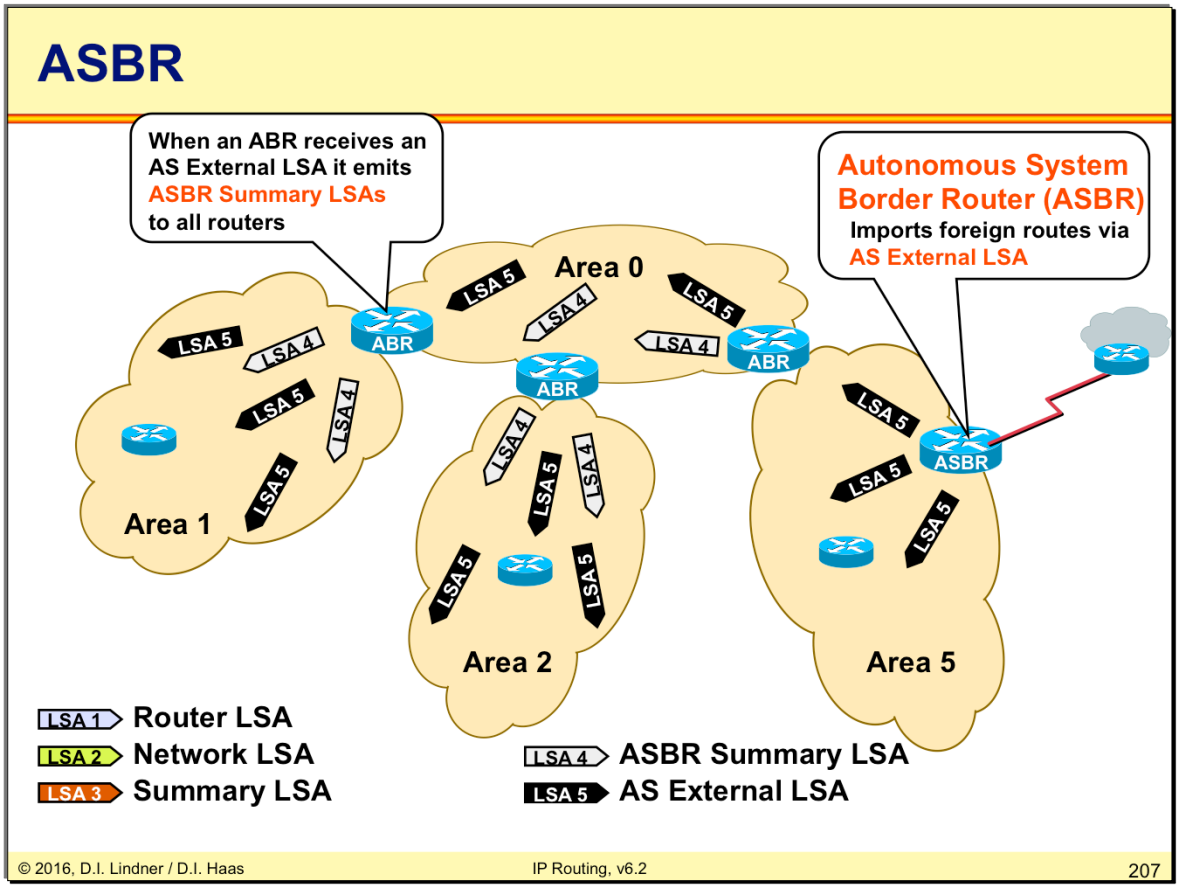
Internal Routers (IR): Has all interfaces inside an area

Backbone Routers (BR): Has at least one interface in the backbone area

Area Border Routers (ABR): Has interfaces in at least two areas

Autonomous System Boundary Routers (ASBR): Has at least one interface in a non-OSPF domain; redistributes external routes into the OSPF domain

ASBRs are discussed next.



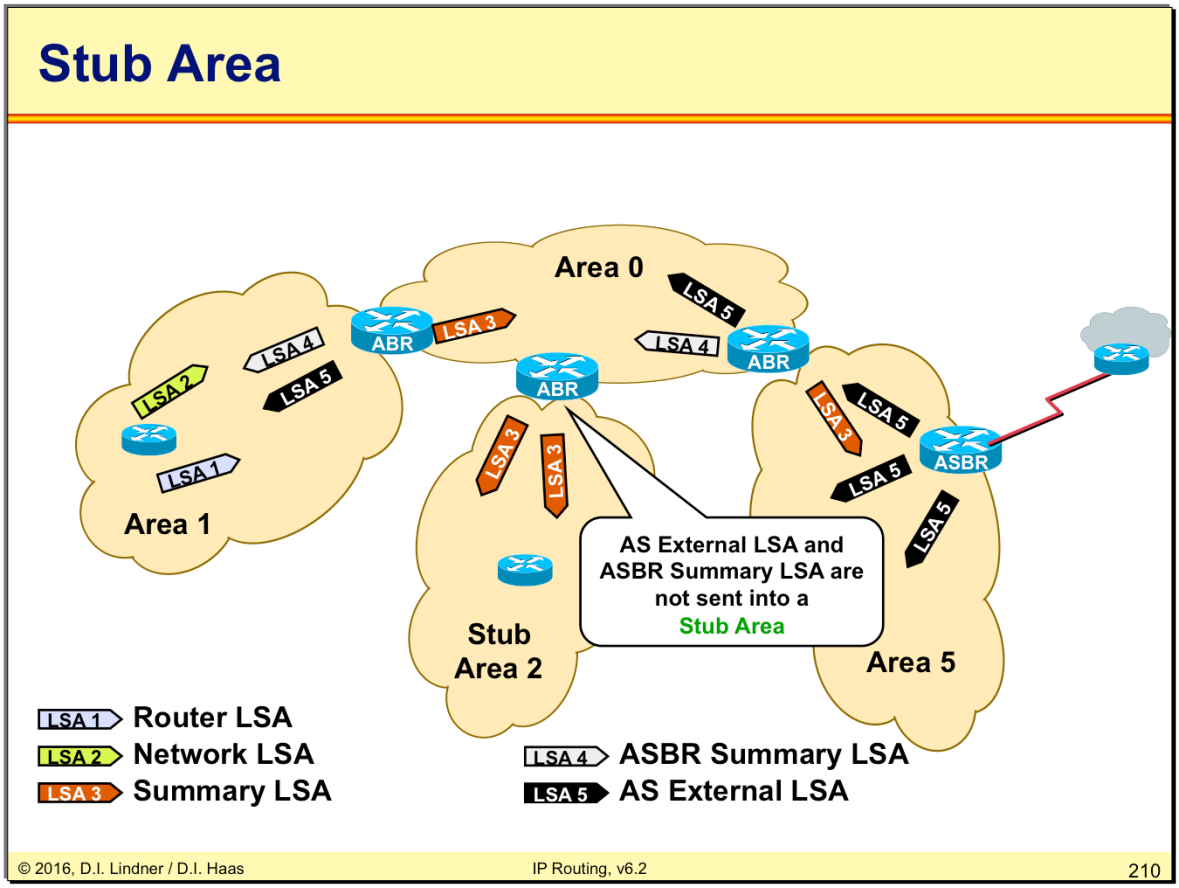
An Autonomous System Border Router (ASBR) sends the summary information about foreign networks to OSPF networks, using LSA type 5. On ASBRs you have to run 2 routing processes: OSPF and some other routing protocol—the router redistributes routing infomation between OSPF and other routing process.

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

OSPF Stub Areas

- **Normally, every internal router gets information about all networks**
  - Internal and external NET-IDs
- **OSPF allows definition of Stub Areas**
  - To minimize memory requirements of internal routers of non-backbone areas for external networks
  - Only the Area Border Router of a particular area knows all external destinations
  - Internal routers only get a default route entry (to this Area Border Router)
  - Any traffic that do not stay inside the OSPF domain (external networks) is forwarded to the Area Border Router

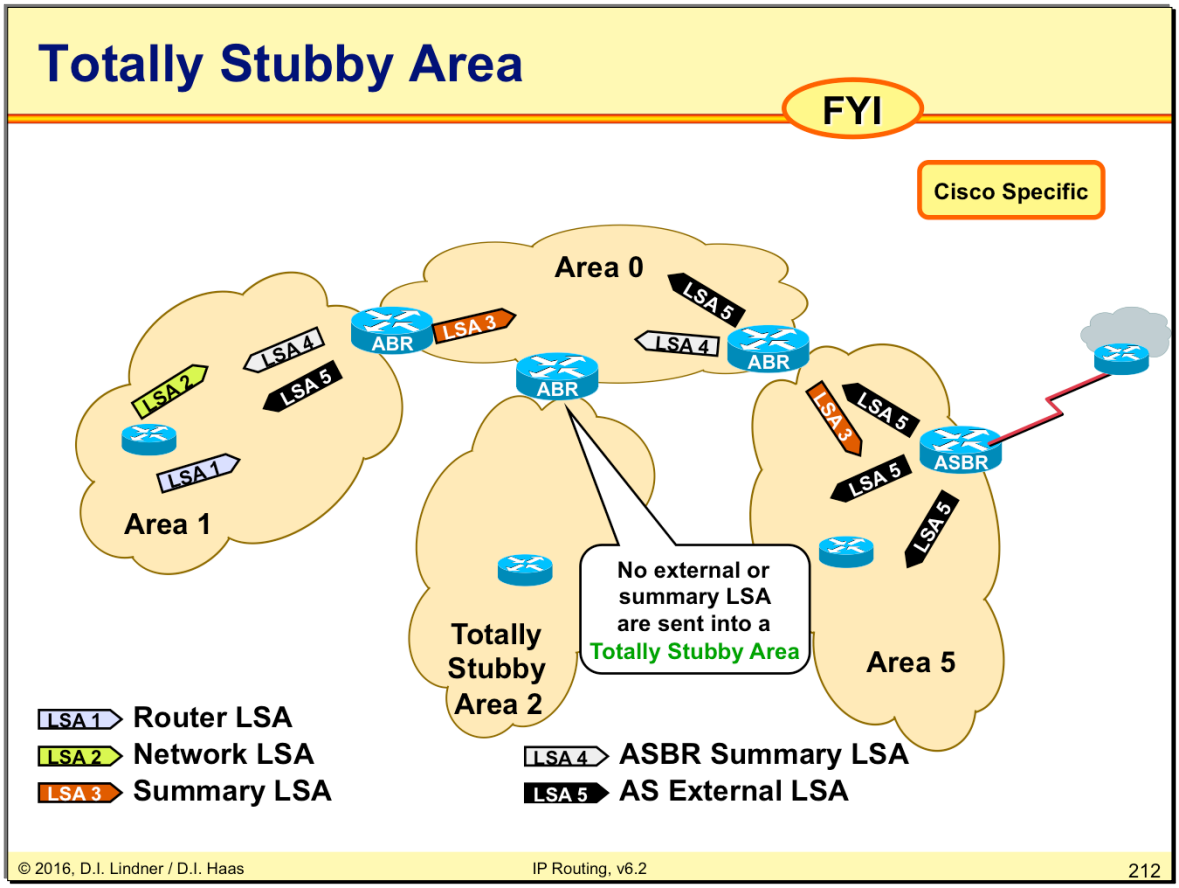


An ASBR could send a lot of external routes, tose will be flooded into OSPF network. ABRs propogate this information into other OSPF areas, each router in the area knows all external links and they are stored in link state database. In order to reach the external destination, the router still needs to send a packet to ABR. We can make a database of internal router smaller, if we create a stub area. A stub area means that ABR does not sent an external LSAs into this area, instead ABR advertises a default route (0.0.0.0)

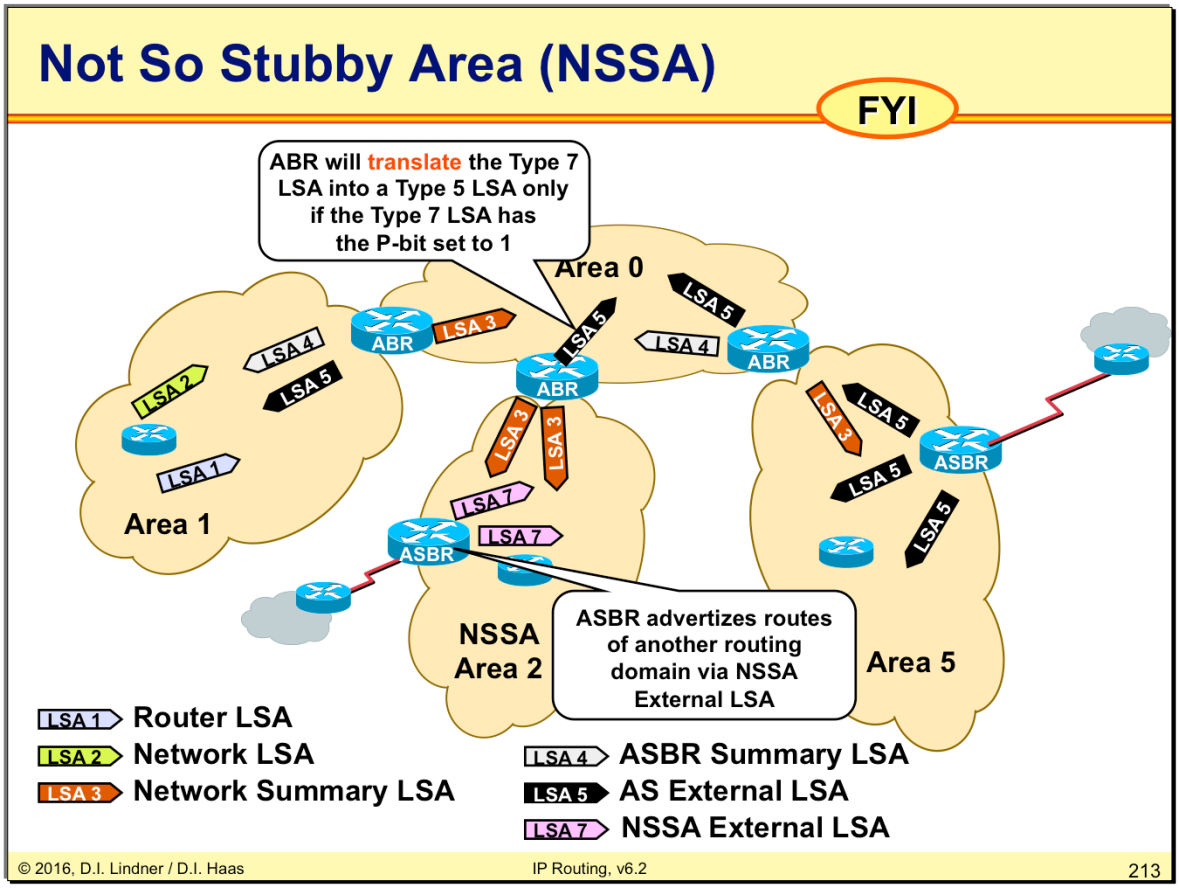
OSPF Totally Stubby Areas

FYI

- **Cisco allows definition of Totally Stubby Areas**
  - Internal routers follow default route also for networks of other areas (no Summary-LSA)
  - That means for internal networks of other areas
- **In such an area**
  - ASBRs are forbidden
- **But if an ASBR should be located in such as totally stubby area**
  - NSSA (Not So Stubby Area) functionality can be used using LSA type 7 updates.



A Cisco‘s proprietary extension to the Stub Area. The ABR will not advertise an external LSAs, like into a stub area, in addition ABR will not send a summary LSAs from other areas, instead a default route is injected into Totally Stubby area.



The NSSA ASBR has the option of setting or clearing the P-bit in the NSSA External LSA. If the P-bit is set any ABR will translate this LSA into an AS External LSA (Type 5).

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Summary LSA and Route Summarization

- **Summary LSA is generated by Area Border Router to inform**
  - Routers inside its area about costs of networks from outside (message direction: Backbone Area -> Area)

    --> import of net-IDs
  - Routers outside its area about costs of its internal networks (message direction: Area -> Backbone Area)

    --> export of net-IDs
- **Additionally Summary Link LSA can be used for Route Summarization**
  - Several net-IDs can be summarized to a single net-ID using an appropriate subnet-mask

Route Summarization 1

- **Route Summarization can be configured manually for Area Border Routers**
  - To minimize number of routing table entries
  - To provide decoupling of OSPF areas
- **Basically, an OSPF domain allows combining any IP-address with any arbitrary subnet masks**
  - Classless Routing
- **No automatic Route Summarization at the IP address class boundary (A,B or C) like RIPv1**
  - Note: RIPv1 implements Classful Routing

Route Summarization 2

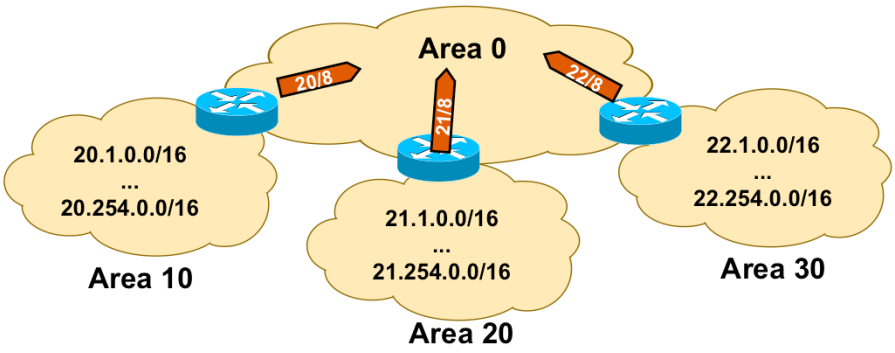
- **Summarization can occur at any place of the IP-address**
- **For instance, many class C addresses can be summarized to one single address (with a prefix)**
  - E.g. class C addresses 201.1.0.0 to 201.1.255.0 (subnet-mask 255.255.255.0) can be summarized by a single entry 201.1.0.0 with subnet-mask 255.255.0.0
  - Note1: when summarizing several networks, only the lowest costs of all these networks are reported (RFC 1583)
  - Note2: when summarizing several networks, only the highest costs of all these networks are reported (RFC 2328)

Route Summarization 3

- **OSPF Route Summarization demands**
  - A clever assignment of IP-addresses and areas to enable Route Summarization
- **Hence OSPF not only forces a star shaped area topology but also demands for a sound IP-address design**
- **Note:**
  - It is still possible to use arbitrary subnet masks and arbitrary addresses anywhere in the network because of classless routing
  - In conflict cases ”Longest Match Routing Rule" is applied
  - But this means a bad network design

Example Summarization

- Efficient OSPF address design requires hierarchical addressing
- Address plan should support summarization at ABRs



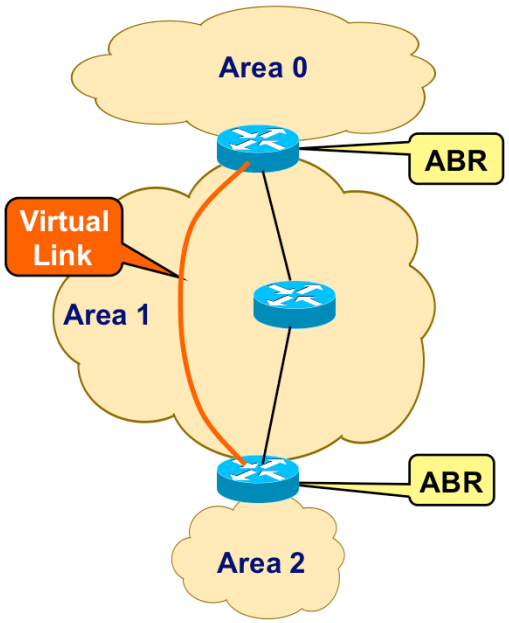
Summarization is an other way to keep a router database smaller. The ABR instead of sending each single subnet from the area, creates a summary route and advertises it into a different area. Note that summarization is turned off by default (i. e. must be explicitly turned on).

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link **FYI**
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Virtual Links

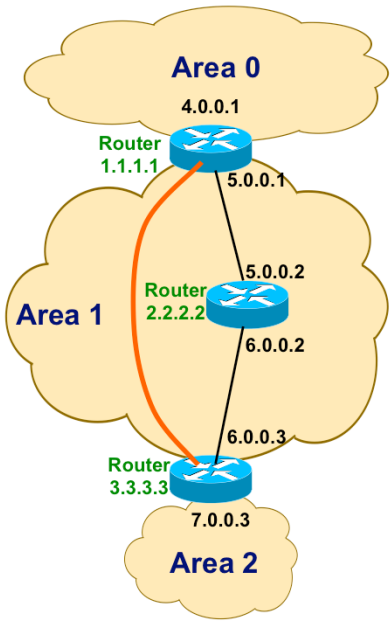
- **Another way to connect to area 0 using a point-to-point tunnel**
- **Transit area must have full routing information**
  - Must *not* be stub area
- **Bad Design!**



An OSPF design requires that all areas have to be contiguous and must be connected to the backbone area. If it is not a case, like on the slide, you have to use a Virtual Link in order to connect area 2 to area 0 A virtual link is considered as part of area 0 thus the area ID is 0.0.0.0.

Virtual Link Example

- **Now router 3.3.3.3 has an interface in area 0**
- **Thus router 3.3.3.3 becomes an ABR**
  - Generates summary LSA for network 7.0.0.0/8 into area 1 and area 0
  - Also summary LSAs in area 2 for all the information it learned from areas 0 and 1



 IP Routing, v6.2 222

A router 3.3.3.3 is now connected to area 0 „directly“ and like a normal ABR generates a summary LSAs in both directions

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details
- **Introduction to Internet Routing (BGP, CIDR)**

Distance-Vector **versus** Link-State

- Distance-Vector:
  - Every router notifies directly connected routers about all reachable routes
  - Using broadcast messages
  - Maintains its routing table according to information from neighbor routers
- Link-State:
  - Every router notifies all routers about the state of his directly connected links
  - Using flooding mechanism (LSA)
  - Calculates optimal paths whenever a new LSA is received

OSPF Benefits 1

- **Network load is significantly smaller than that of distance vector protocols**
  - Short hello messages between adjacent routers versus periodical emission of the whole routing table
- **Even update messages after topology modifications are smaller than the routing table of distance vector protocols**
  - LSAs only describe the local links for which a router is responsible -> incremental updates !!!
- **Massive network load**
  - Occurs only on combining large splitted network parts of an OSPF domain (many database synchronizations)

OSPF Benefits 2

- **SPF-techniques take advantages from several features:**
  - Every router maintains a complete topology-map of the entire network and calculates independently its desired paths (actually based on the original LSA message)
  - This local ability for route calculation grants a fast convergence
  - LSA is not modified by intermediate routers across the network
  - The size of LSAs depends on the number of direct links of a router to other routers and not on the number of subnets!

OSPF Benefits 3

- **During router configuration, every physical port is assigned a cost value**
  - Per ToS (Type of Service)
  - Each ToS can be assigned a separate topology map (8 possible combinations)
  - IP's ToS field may examined for packet forwarding
    - Note: OSPF ToS support disappeared in RFC 2328
- **Determination of the best path for a specific ToS is based on the summary costs along the paths**
  - RIP uses hop count only
- **Equal costs automatically enables load balancing between these paths**

OSPF Benefits 4

- **Subnet masks of variable length can be attached to routes (in contrast to RIPv1)**
- **External routes are marked (tagged) explicitly to be differentiated from internal routes**
- **OSPF messages can be authenticated to grant secure update information**
- **OSPF routing messages use IP-multicast addresses: lower processing effort**
- **Point-to-point connections do not need own IP-address**
  - In theory more economic use of address space is possible
  - But for practical reasons regarding network management also on point-to-point connections usage of IP addresses are recommended

OSPF in Large Networks

- **OSPF area concept can be used**
  - A two level hierarchy is used to decrease
    - CPU time for SPF calculations
    - Memory requirement for storing topology database
  - One backbone area
  - Several non-backbone areas
    - Non-backbone area can be connected by area border router to backbone area only
  - Summarization possible at area border routers
    - Route aggregation to reduce size of routing tables
    - Summarization means that some net-IDs can be summarized as one net-ID only

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
  - Introduction
  - The Dijkstra Algorithm
  - Communication Procedures
  - LSA Broadcast Handling
  - Split Area
  - Broadcast Networks
  - Area Principles
  - Stub Areas
  - Route Summarization
  - Virtual Link
  - Summary
  - OSPF Header Details FYI
- **Introduction to Internet Routing (BGP, CIDR)**

Router LSA – Type 1

- **Router ID (Highest IP address)**
- **Number of Links**
- **Link Descriptions**
  - Link type (P2P, Stub, ...)
  - Neighboring router ID
  - Router interface address
  - ToS (typically not supported today)
  - Metrics

Network LSA – Type 2

- DR's IP address
- One Subnet mask for this broadcast segment
- List of Router-IDs of all routers in the broadcast segment

Network Summary LSA – Type 3

- **Originated by ABRs only**
- **Each LSA Type 3 contains a number of**
  - Destination networks + Subnet masks
  - Metric for each destination network
- **This is basically a distance-vector routing information (!)**

ASBR Summary LSA – Type 4

- **Originated by ABRs**
- **Advertise routes to ASBRs**
- **Nearly identical to Type 3**
  - Except destination is ASBR not a network
- **Each LSA Type 4 contains**
  - Router IDs of ASBRs
  - Mask 0.0.0.0 (host route)
  - Metric

AS External LSA – Type 5

- **Originated by ASBRs**
  - External type 1
  - External type 2 (default)
- **Advertises**
  - External routes
  - Default route
- **Contains**
  - External Net-ID + Mask
  - Metric
  - Next hop (external, not ASBR)

NSSA External LSA – Type 7

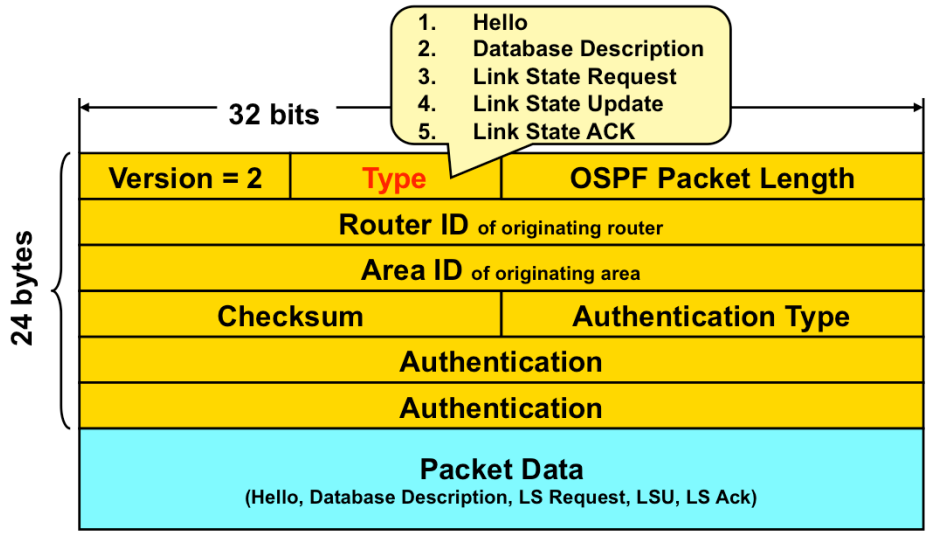
- **Originated by ASBRs within NSSAs**
- **Almost identical to Type 5**
  - But only flooded within NSSA
- **RFC 1587**

Other LSAs

- **Group Membership LSA (6)**
  - For MOSPF
- **External Attribute LSA (8)**
  - Alternative to IBGP
  - Should transport BGP information within an OSPF domain
  - Not yet implemented, no RFC yet (?)
- **Opaque LSA (9)**
  - Application specific information
  - Link local scope
- **Opaque LSA (10)**
  - Application specific information
  - Area-local scope
- **Opaque LSA (11)**
  - Application specific information
  - AS scope

Opaque LSAs are e. g. used as load indication messages with MPLS.

General OSPF Packet Structure



- Carried directly in IP (protocol number 89)
- All OSPF packets begin with a 24-byte OSPF packet header

 IP Routing, v6.2 238

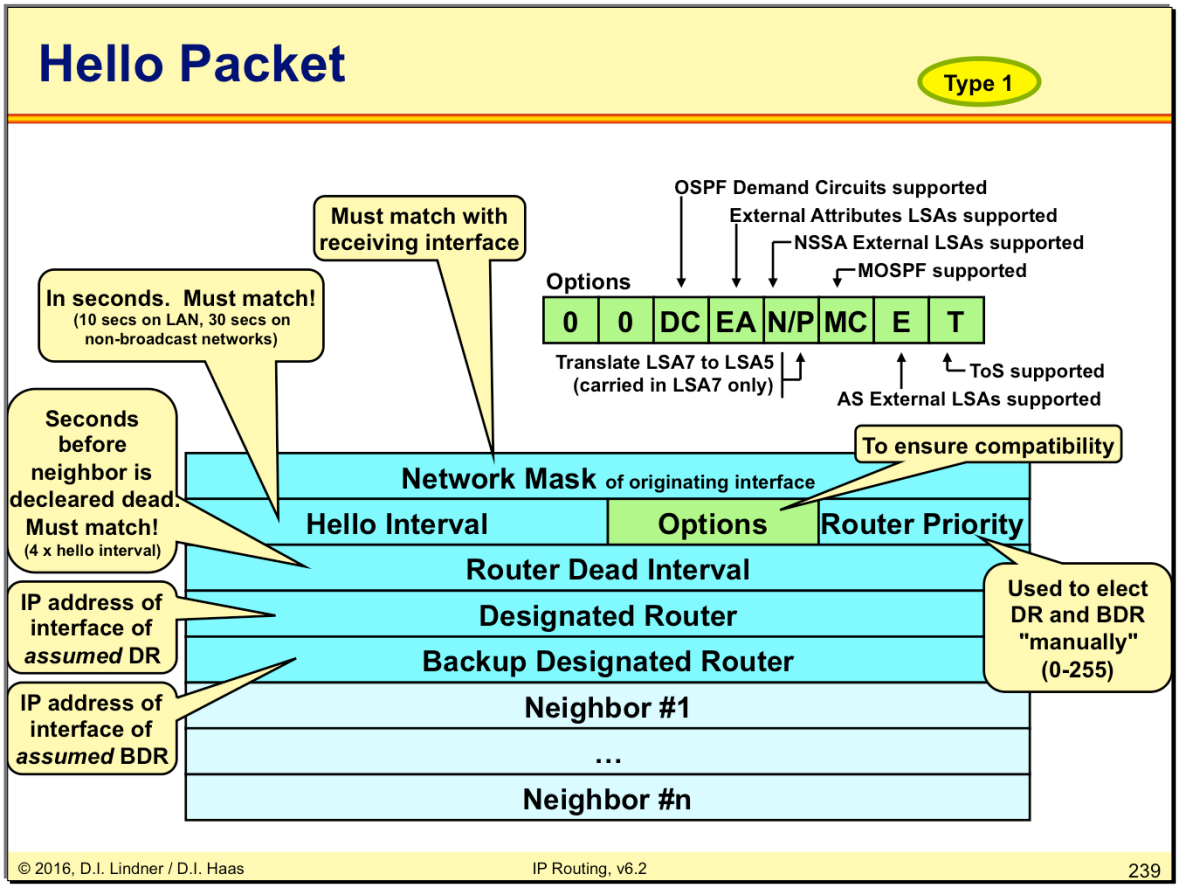
The OSPF version we use today is version 2. The packet type identifies the actual OSPF message type that is carried in the packet data area at the bottom. The OSPF packet length describes the number of bytes of the OSPF packet including the OSPF header. Router and Area IDs identify the originator of this packet. If a packet is sent over a virtual link, the Area ID will be 0.0.0.0, because virtual links are considered part of the backbone area. The checksum is calculates over the entire packet including the header.

Three authentication types had been defined:

0	No authentication
1	Simple clear text password authentication
2	MD5 Checksum

If the Authentication Type = 1, then a 64 bit clear text password is carried in the authentication fields. If the Authentication Type = 2, then the authentication fields contain a key-ID, the length of the message digest, and a non decreasing cryptographic sequence number to prevent replay attacks. The actual message digest would be appended at the end of the packet.

The efficiency of routing updates also depends on the maximum transfer unit (MTU) defined. Cisco defined a MTU of 1500 bytes for OSPF.



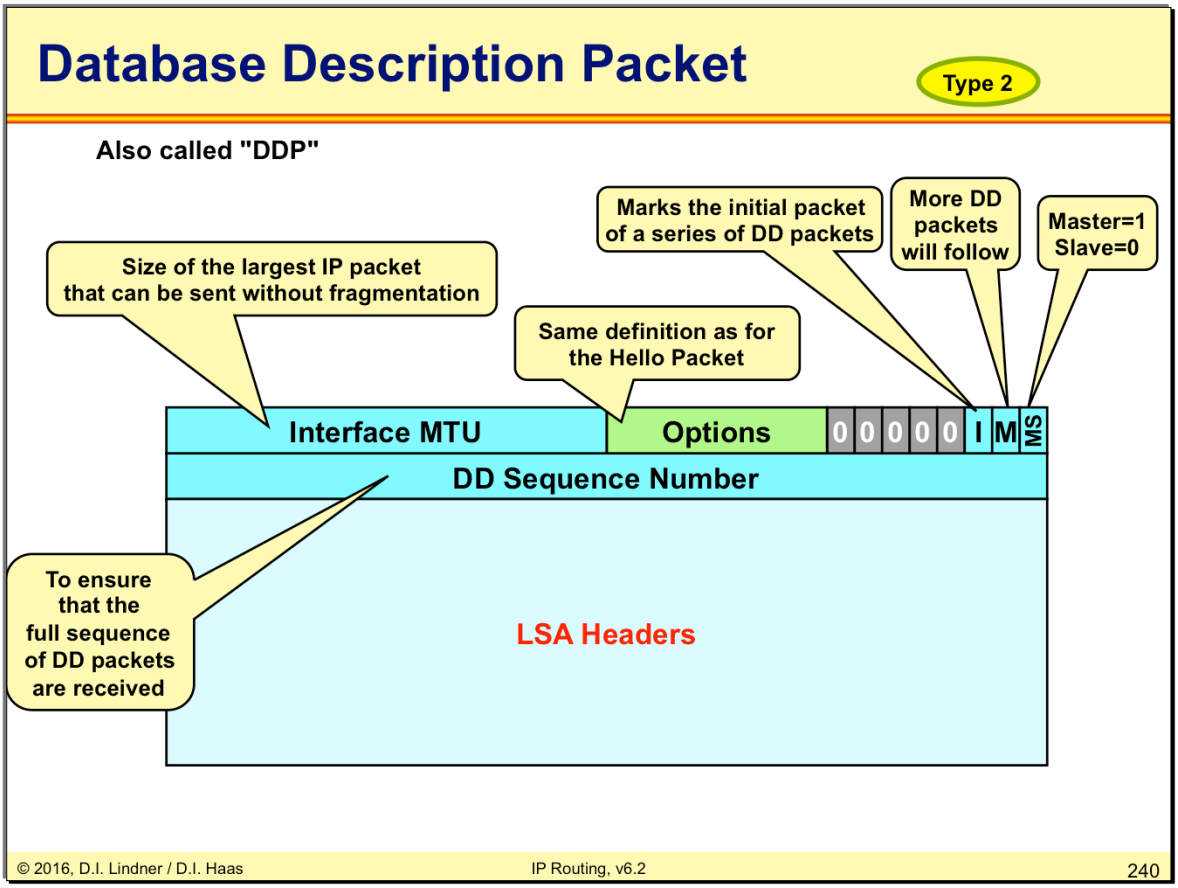
The network mask must match the mask on the receiving interface, ensuring that they share a segment and network.

The Options field is also used by other message types. If the Router Priority is set to zero this router cannot become DR or BDR.

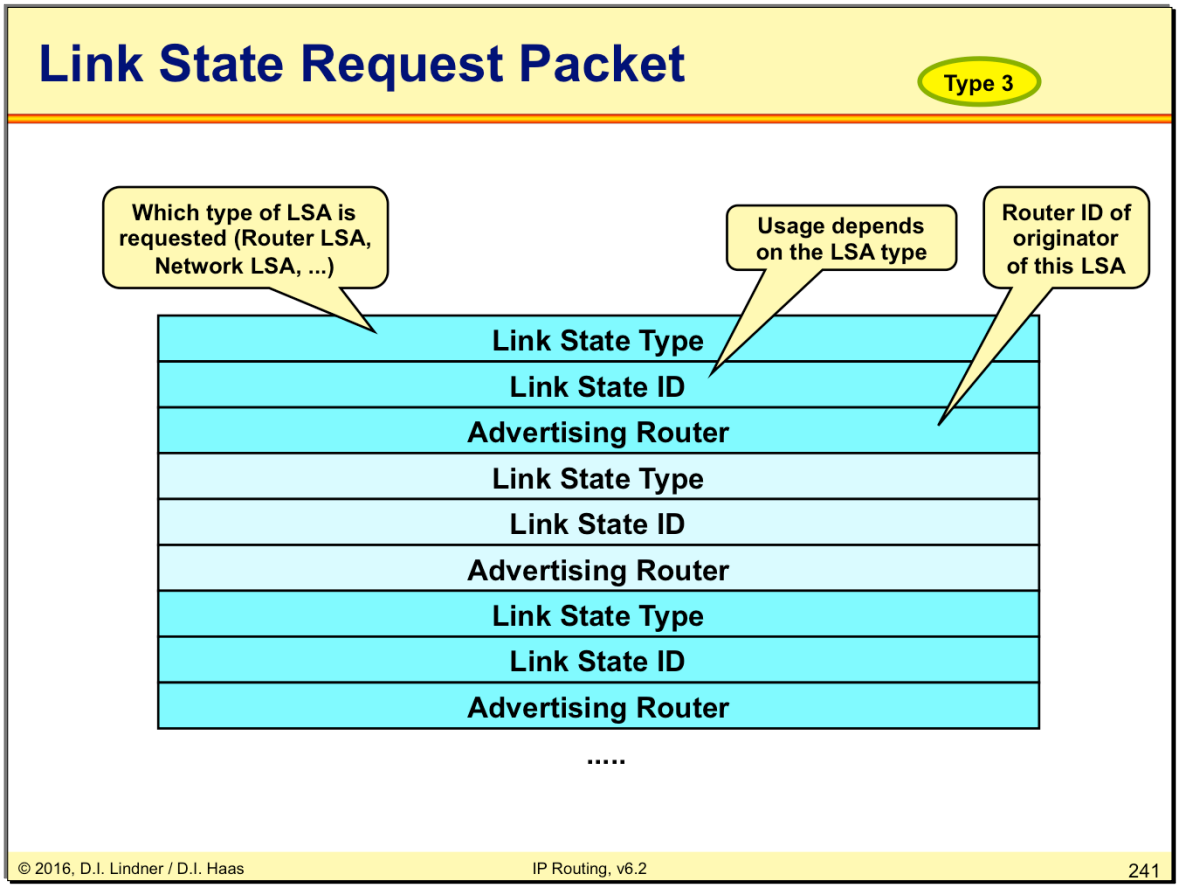
Note that the fields "Designated Router" and "Backup Designated Router" only contain the interface IP address of the DR or BDR on that network, not the router ID !!

If these numbers are unknown or not necessary (other network type) then these fields are set to 0.0.0.0.

It is important to know that neighbors must have configured identical Hello and Dead Intervals.



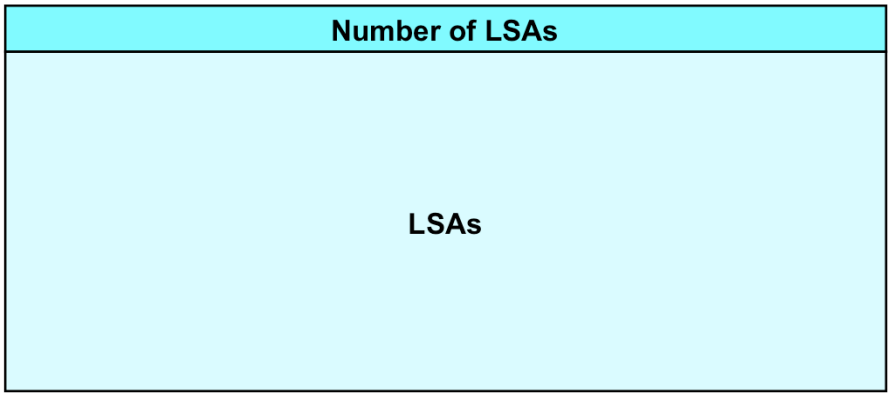
The DD sequence number is set by the master to some unique value in the first DD packet. This number will be incremented in subsequent packets.



Note that the Link State Request Packet uniquely identifies the LSA by Type, ID, and advertising router fields of its header. It does not include the sequence number, checksum, and age, because the requestor is not interested in a specific instance of the LSA but in the most recent instance.

Link State Update Packet

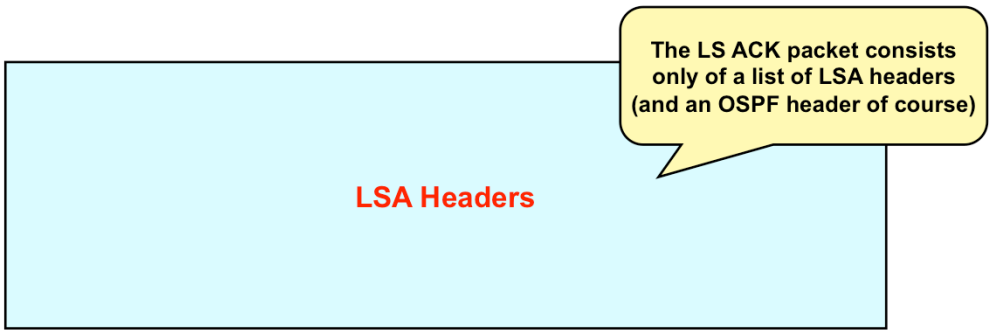
Type 4



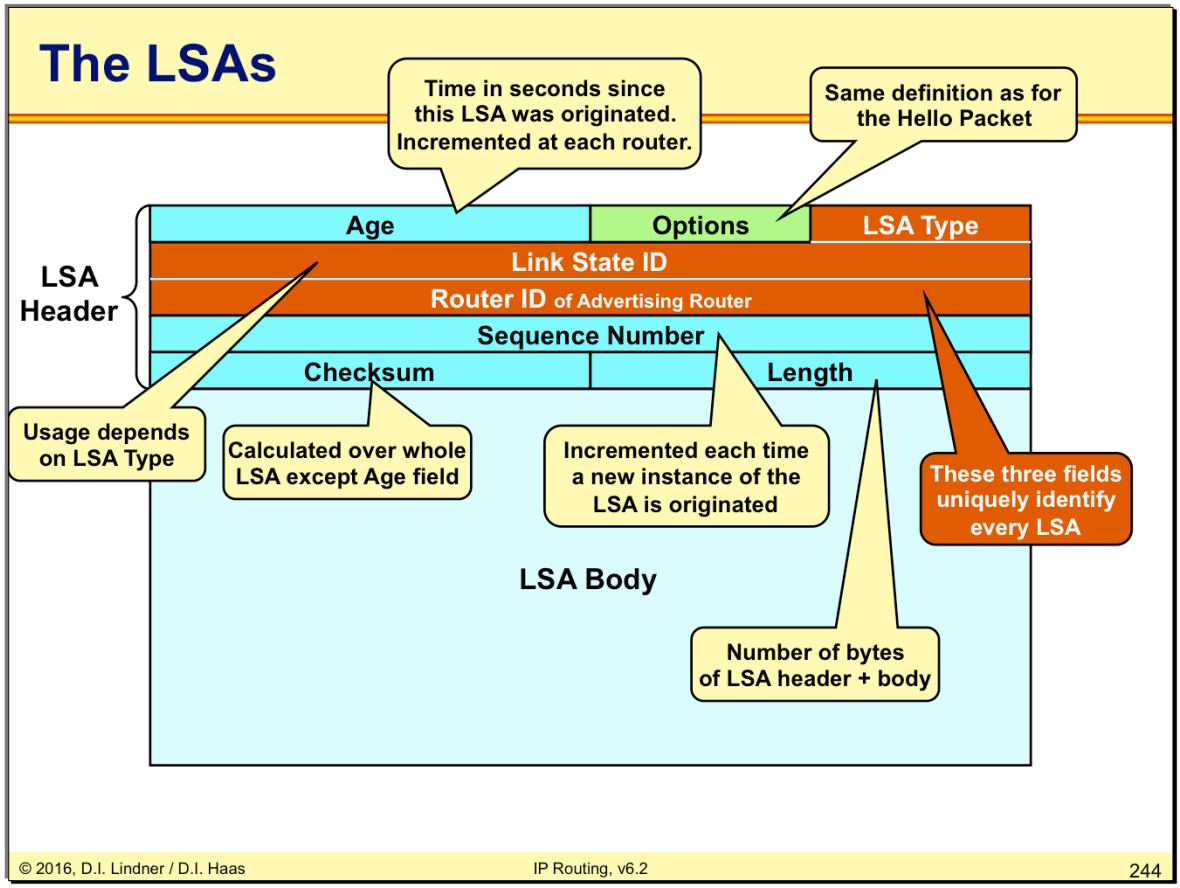
- LSUs contain one or more LSAs (limited by MTU)
- Used for flooding and response to LS requests
- LSUs are carried hop-by-hop

Link State ACK Packet

Type 5



- Each LSA received must be explicitely acknowledged ➔ reliable flooding!
- Acknowledged LSA is identified by LSA header
- Single Link State ACK packet can acknowledge multiple LSAs



All LSAs have the LSA header at the beginning. This LSA header is also used in Database Description and Link State Acknowledgement packets.

The Age is incremented by InfTransDelay seconds at each router interface this LSA exits. The Age is also incremented in seconds as it resides in a link state database.

The Options field describes optional capabilities supported at that topological portion described by this LSA.

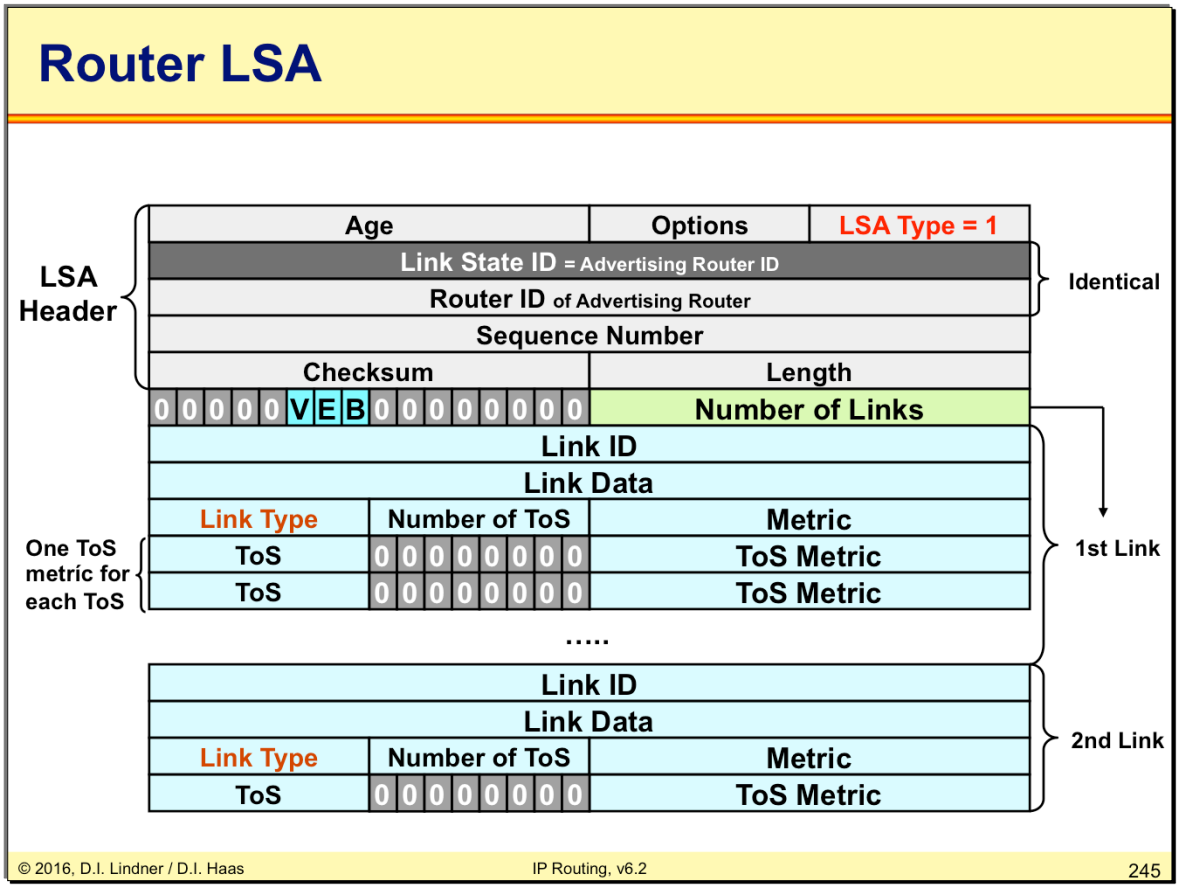
The LSA Type describes which information is carried in the LSA Body. Here the structural differences between Router LSAs, Network LSAs, etc. are identifed.

The Link State ID is used differently by the LSA types. Basically this field contains some information identifying the topological portion described by this LSA. For example a Router ID or an interface address is used here. The following slides will explain this field for each LSA type.

The Router ID identifies the originating router of this LSA.

The Sequence Number helps routers to identify the most recent instance of this LSA.

The Checksum is a so-called 8 bit Fletcher checksum, providing more protection than traditional checksum methods such as used for TCP. The first eight bits contain the 1's complement sum of all octets, while the second eight bits contain a high-order sum of the running sums. See RFC 1146 for more details.



Router LSAs are generated by all OSPF routers and must describe all links of the originating router!

The V-bit (Virtual Link Endpoint) is set to one if the originating router is a virtual link endpoint and this area is a transit area. The E-bit (External) is set if the originating router is an ASBR. The B-bit (Border) is set if the originating router is an ABR.

The Link ID and Link Data depend on the Link Type field which describes the general type of connection the link provides.

Link Type 1 is a point-to-point link, the Link ID describes the Neighbor Router ID and the Link Data field contains the IP address of the originating router's interface to the network.

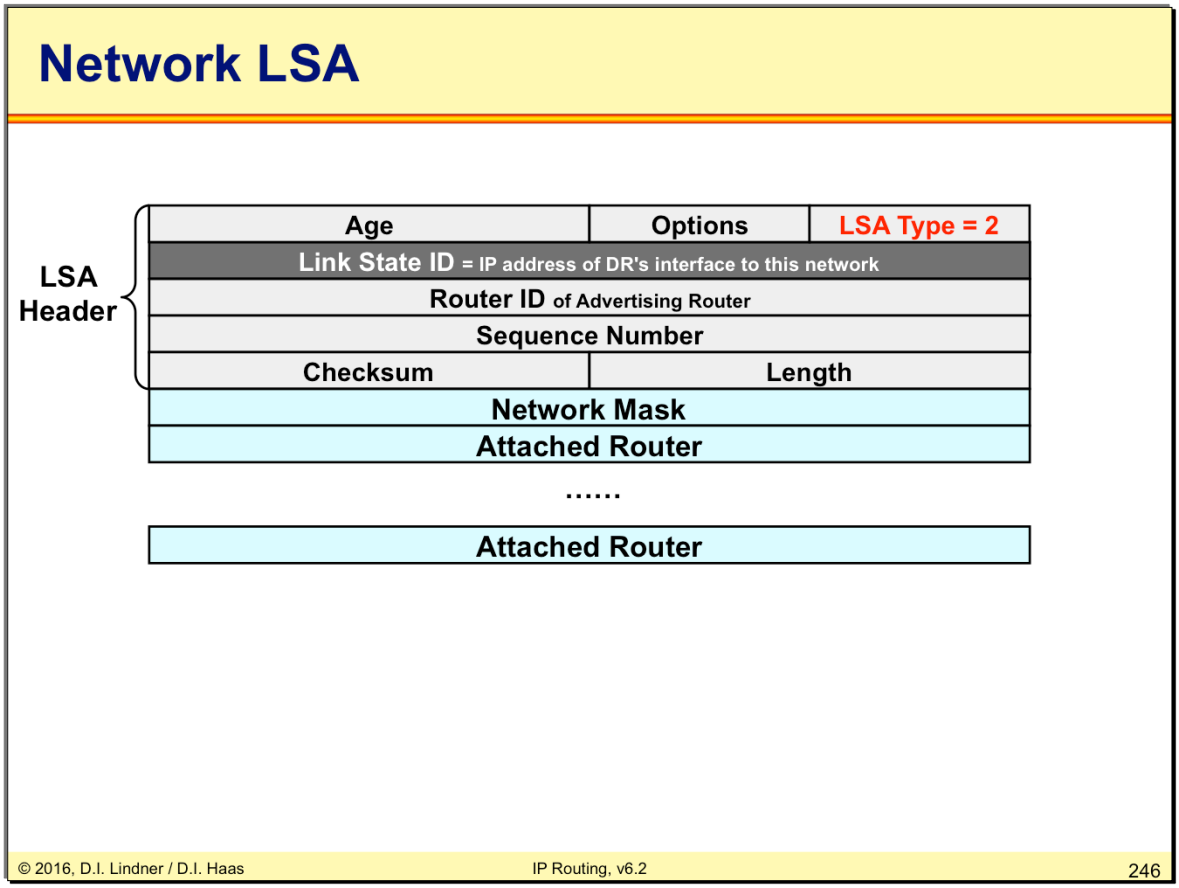
Link Type 2 is a link to a transit network, the Link ID describes the interface address of the Designated Router and the Link Data field contains the IP address of the originating router's interface to the network.

Link Type 3 is a link to stub network, the Link ID describes the IP network number or subnet address and the Link Data field contains the network's IP address or subnet mask.

Link Type 4 is a virtual link, the Link ID describes the neighboring router's Router ID and the Link Data contains the MIB-II ifIndex value for the originating router's interface.

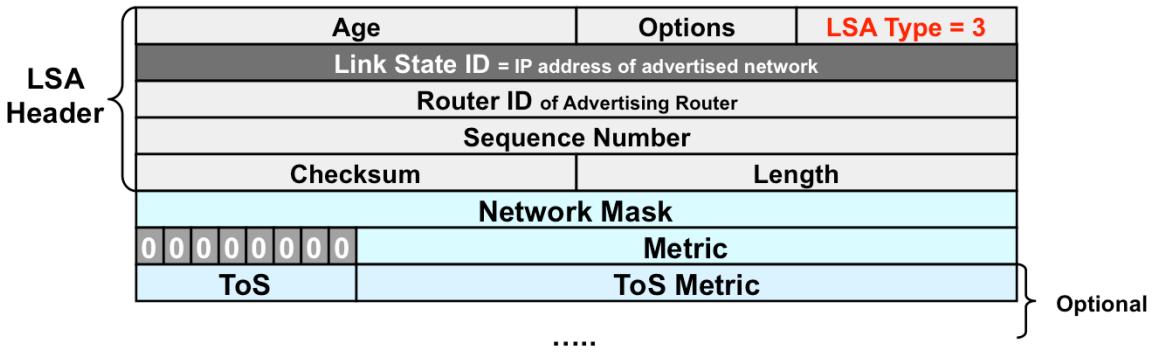
Number of ToS specifies the number of ToS Metrics listed for this link. For each ToS an additional line is appended to this link state section. Generally, ToS is not used today anymore and the Number of ToS field is set to all-zero.

Metric is the cost of the interface that established this link.



Network LSAs are originated by DRs and describe the multi-access network and all routers attached to it, including the DR.

Network Summary LSA

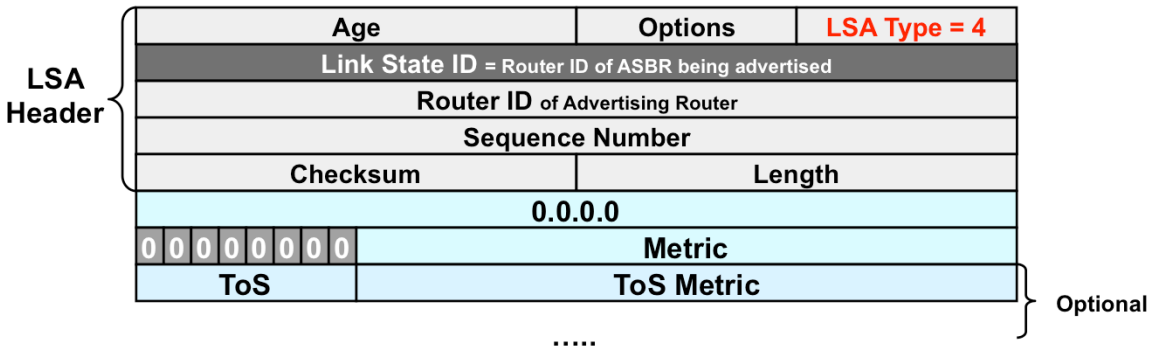


- If a default route is advertised, both the Link State ID and the Network Mask fields will be 0.0.0.0
- Also used for route summarization
- Note: Cisco only supports ToS=0

 IP Routing, v6.2 247

A Network Summary LSA is originated by an ABR and advertises networks external to an area.

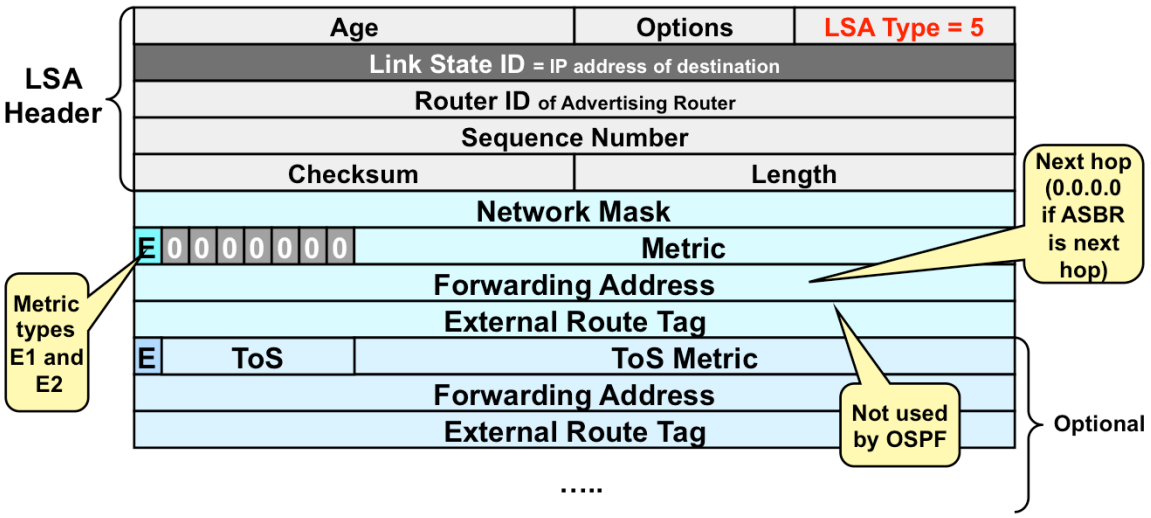
ASBR Summary LSA



- Note: Cisco only supports ToS=0

A ASBR Summary LSA is originated by an ABR and advertises ASBRs external to an area.

Autonomous System External LSA



- **When describing a default route, both the Link State ID and the Network Mask are set to 0.0.0.0.**

NSSA External LSA

- **Same structure as AS External LSA**
- **Forwarding address is**
  - Next hop address for the network between NSSA and adjacent AS, if this network is advertised as internal route
  - Router ID of NSSA-ASBR otherwise

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
  - Introduction
  - BGP Basics
  - BGP Attributes
  - BGP Special Topics
  - CIDR

Routing in Small Networks

- **In small networks**
  - Distance vector or link state protocols like RIP or OSPF can be used for dynamic routing
  - It is possible that every router of the network knows about all destinations
    - All destination networks will appear in the routing tables
  - Routing decisions are based on technical parameters
    - E.g. hop count, link bandwidth, link delay, interface costs
  - It is sufficient that routing relies only on technical parameters
    - Small networks will be administered by a single authority
    - Non-technical parameter like traffic contracts have no importance

Routing in Large Networks

- **With increasing network size limitations of these protocols can be recognized**
  - Some limitations for example
    - Maximum hop count (RIP)
    - Time to transmit routing tables (RIP) on low speed links
    - CPU time for SPF calculation (OSPF)
    - Memory used for storing routing table (RIP, OSPF)
    - Memory used for storing topology database (OSPF)
    - Two level hierarchy centered around a core network (OSPF)
    - Route fluctuation caused by link instabilities (OSPF)
    - Routing based on non-technical criteria like financial contracts or legal rules is not possible

Routing in the Internet

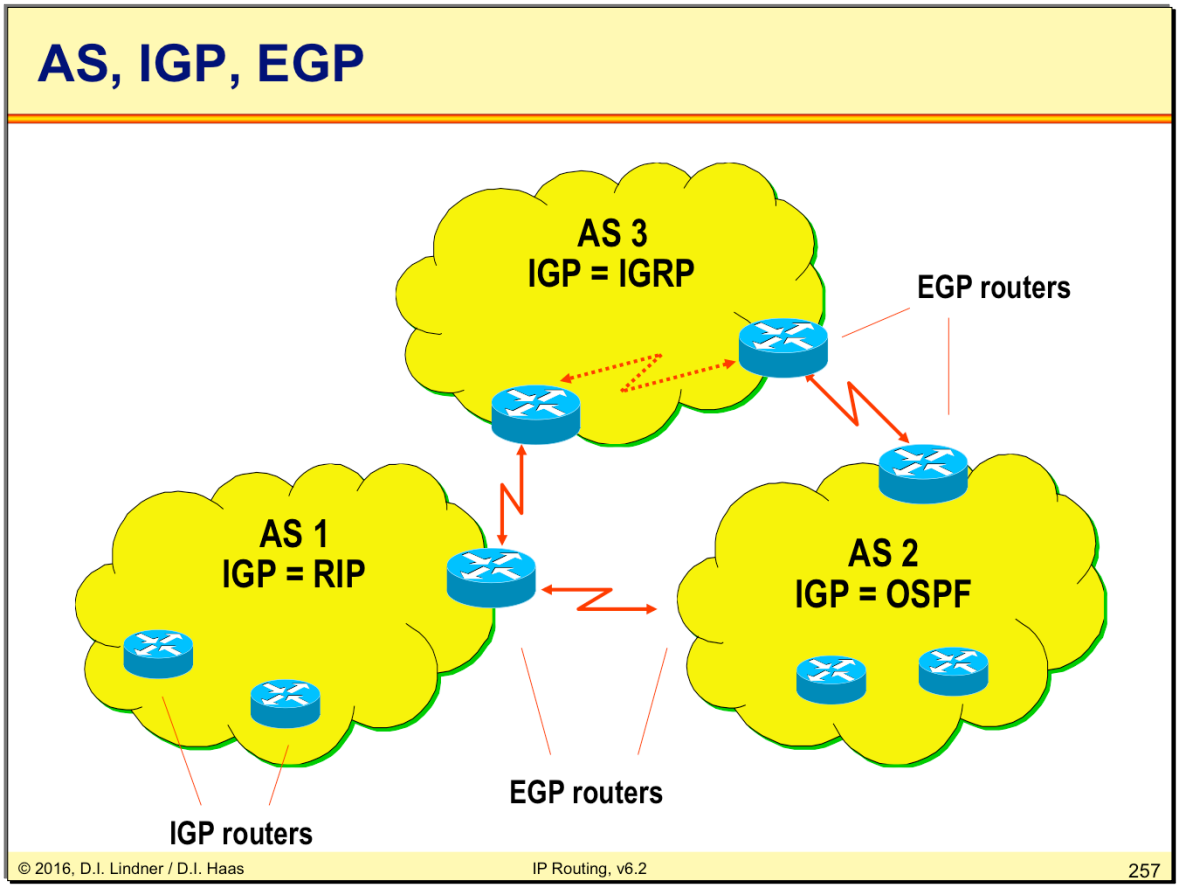
- **Limitations prevent using routing protocols like RIP or OSPF for routing in the Internet**
  - Note: routing tables of Internet-core routers have about 415.000 net-ID entries (May 2012)
- **Routing in the Internet**
  - Is based on non-technical criteria like financial contracts or legal rules
  - Policy routing
    - Acceptable Use Policy (AUP) in parts of the Internet
    - Contracts between Internet Service Providers (ISP)

Routing Hierarchy, Autonomous Systems

- **Routing hierarchy is necessary for large networks**
  - To control expansion of routing tables
  - To provide a more structured view of the Internet
- **Routing hierarchy used in the Internet**
  - Based on concept of autonomous system (AS)
- **AS concept allows**
  - Segregation of routing domains into separate administrations
  - Note: routing domain is a set of networks and routers having a single routing policy running under a single administration

IGP, EGP

- **Within an AS one or more IGP protocols provide interior routing**
  - IGP - Interior Gateway Protocol
  - IGP examples
    - RIP, RIPv2, OSPF, IGRP, eIGRP, Integrated IS-IS
  - IGP router responsible for routing to internal destinations
- **Routing information between ASs is exchanged via EGP protocols**
  - EGP - Exterior Gateway Protocols
    - EGP router knows how to reach destination networks of other ASs
  - EGP examples
    - EGP-2, BGP-3, BGP-4


AS, IGP, EGP
AS 3
IGP = IGRP
EGP routers
AS 1
IGP = RIP
AS 2
IGP = OSPF
EGP routers
IGP routers
© 2016, D.I. Lindner / D.I. Haas
IP Routing, v6.2
257

AS Numbers

- **Hierarchy based on ASs allows forming of a large network**
  - By dividing it into smaller and more manageable units
  - Every unit may have its own set of rules and policies
- **AS are identified by a unique number**
  - Can be obtained like IP address from an Internet Registry
    - e.g. RIPE NCC (Reséaux IP Européens Network Coordination Center)

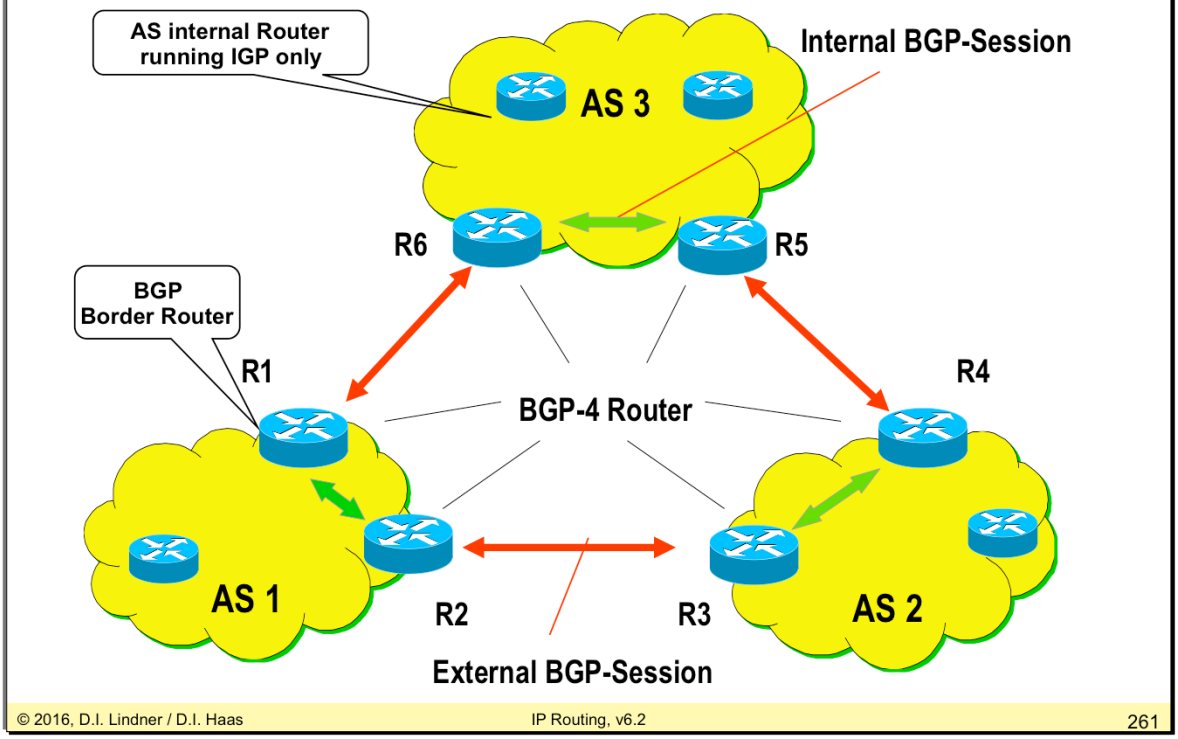
BGP-4 (1)

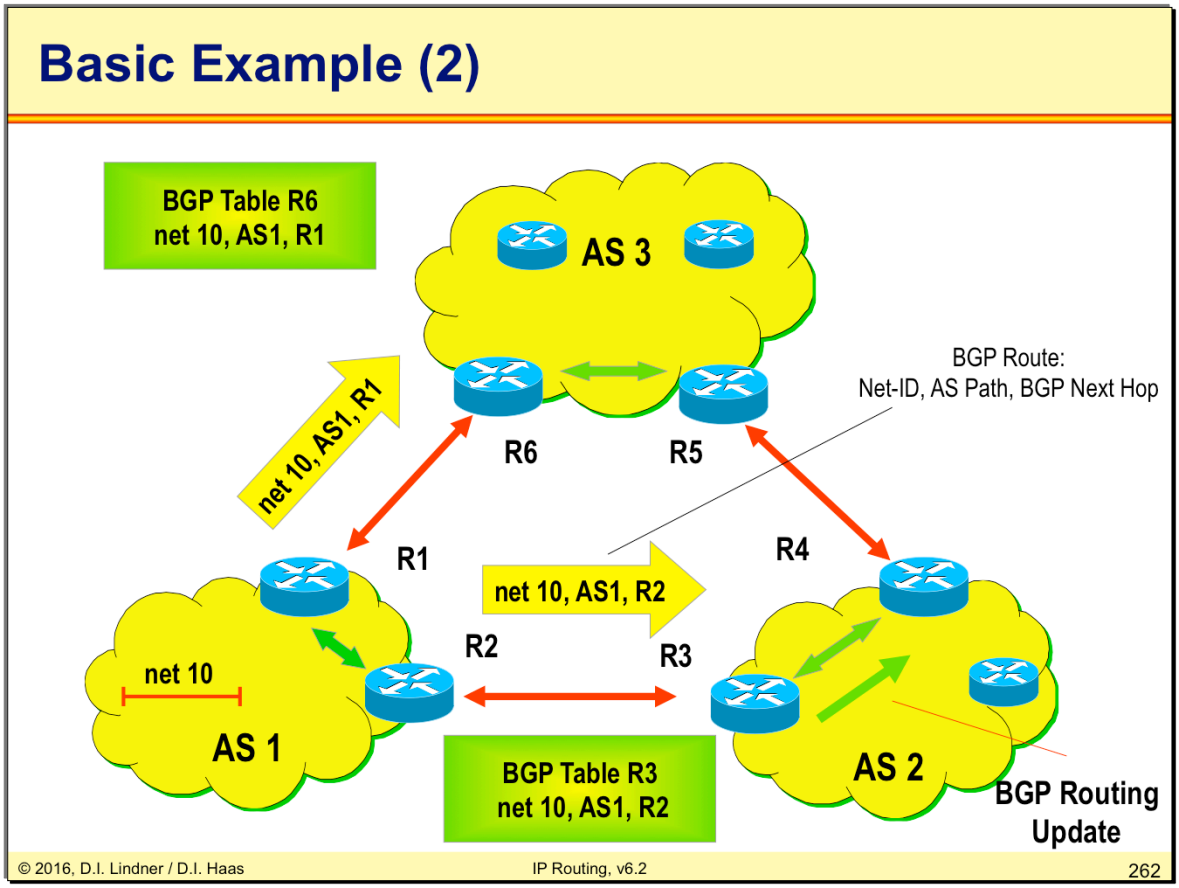
- **Border Gateway Protocol (BGP)**
  - Is the Exterior Gateway Protocol used in the Internet nowadays
  - Was developed to overcome limitations of EGP-2
  - RFC 1267 (BGP-3) older version
    - classful routing only
  - RFC 1771 (BGP-4) current version, DS
    - classless routing
  - Is based on relationship between neighboring BGP-routers
    - Peer to peer
    - Called BGP session or BGP connection

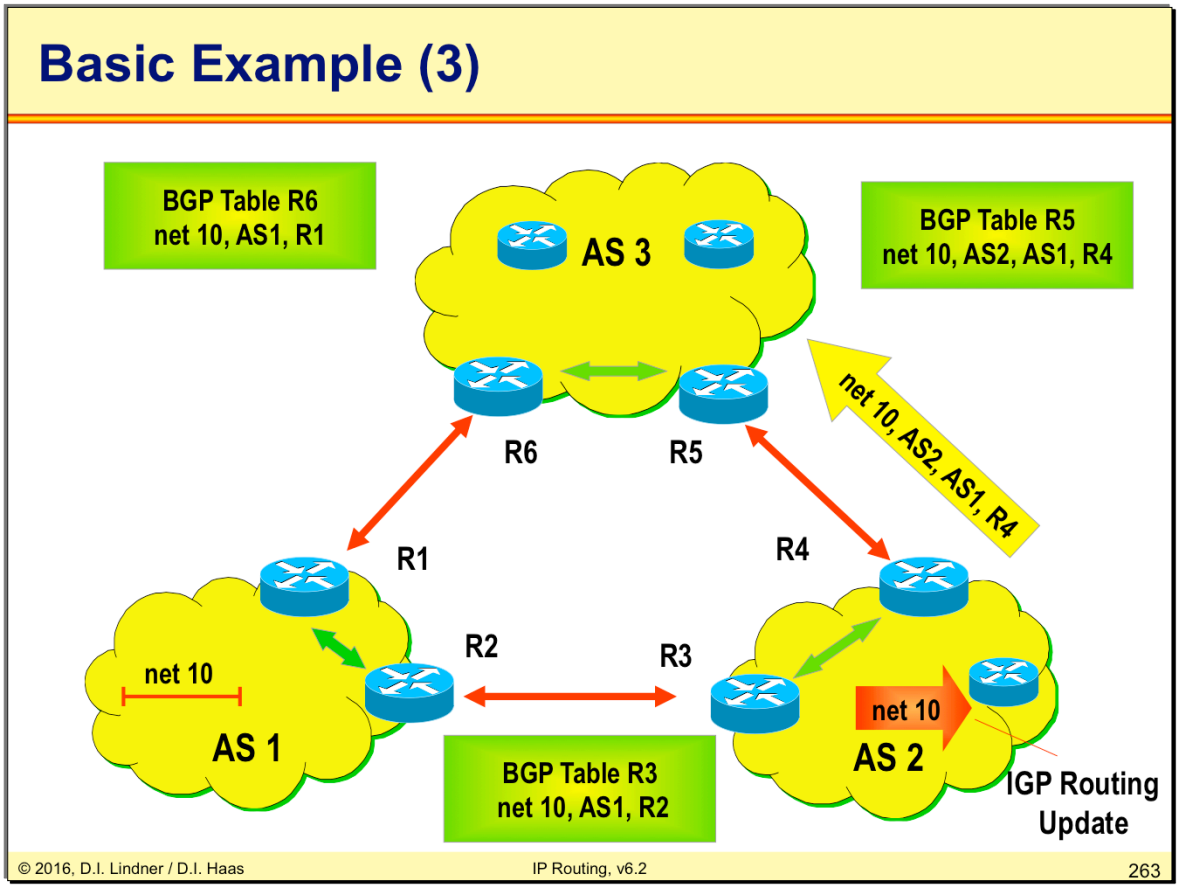
BGP-4 (2)

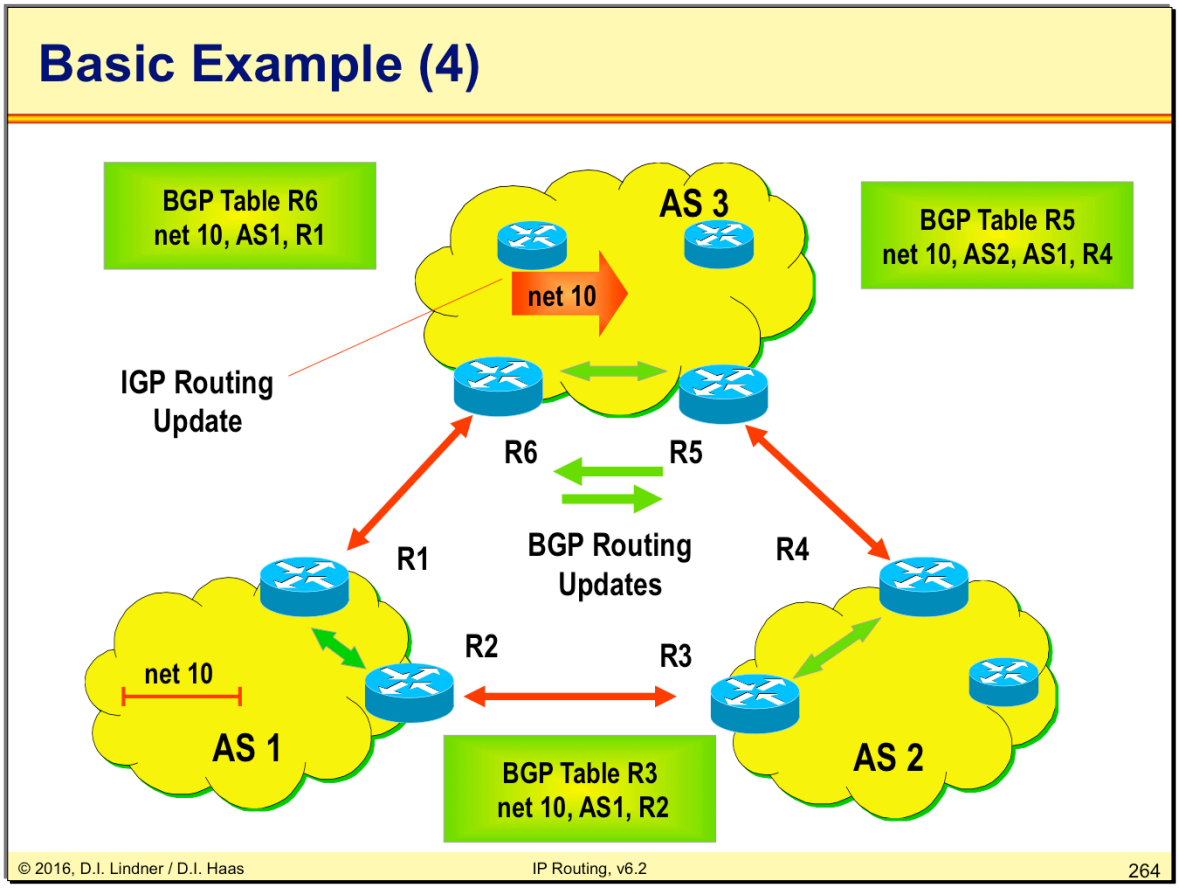
- **Border Gateway Protocol (cont.)**
  - Primary function
    - Exchange of network reachability information with other autonomous systems via external BGP sessions
    - But also within an autonomous system between BGP border routers via internal BGP sessions
  - BGP session runs on top of TCP
    - Reliable transport connection
    - Well known port 179
    - TCP takes care of fragmentation, sequencing, acknowledgement and retransmission
    - Hence these procedures need not be done by the BGP protocol itself

Basic Example (1)




Basic Example (2)
BGP Table R6
net 10, AS1, R1
AS 3
BGP Route:
Net-ID, AS Path, BGP Next Hop
net 10, AS1, R1
R6
R5
R1
R4
net 10, AS1, R2
R2
R3
net 10
AS 1
AS 2
BGP Table R3
net 10, AS1, R2
BGP Routing
Update
© 2016, D.I. Lindner / D.I. Haas
IP Routing, v6.2
262


Basic Example (3)
BGP Table R6
net 10, AS1, R1
AS 3
BGP Table R5
net 10, AS2, AS1, R4
net 10, AS2, AS1, R4
R6
R5
R1
R4
R2
R3
net 10
AS 1
net 10
AS 2
BGP Table R3
net 10, AS1, R2
IGP Routing
Update
© 2016, D.I. Lindner / D.I. Haas
IP Routing, v6.2
263


Basic Example (4)
BGP Table R6
net 10, AS1, R1
AS 3
net 10
BGP Table R5
net 10, AS2, AS1, R4
IGP Routing
Update
R6
R5
BGP Routing
Updates
R1
R4
R2
R3
net 10
AS 1
AS 2
BGP Table R3
net 10, AS1, R2
© 2016, D.I. Lindner / D.I. Haas
IP Routing, v6.2
264

BGP-4 Concepts (1)

- Reachability information exchanged between BGP routers carries a sequence of AS numbers
  - Indicates the path of ASs a route has traversed
- Path vector protocol
- This allows BGP to construct a graph of autonomous systems
  - Loop prevention
  - No restriction on the underlying topology
- The best path
  - Minimum number of AS hops
  - Note: criteria if no other BGP policies are applied
- Incremental update
  - After first full exchange of reachability information between BGP routers only changes are reported

BGP-4 Concepts (2)

- Description of reachability information by BGP attributes
  - For BGP routing
  - For establishing of routing policy between ASs
- BGP-4 advertises so called BGP routes
  - BGP route is unit of information that pairs a destination with the path attributes to that destination
  - AS Path is one among many other BGP attributes
- IP prefix and mask notation
  - Supports VLSM
  - Supports aggregation (CIDR) and supernetting
- Routes can be filtered using attributes, attributes can be manipulated

  --> Routing policy can be established

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
  - Introduction
  - BGP Basics
  - BGP Attributes
  - BGP Special Topics
  - CIDR

Border Gateway Protocol (BGP)

- **BGP-3**
  - Was classful
  - Central AS needed (didn't scale well)
  - Not further discussed here!
  - RFC 1267
- **BGP-4**
  - Classless
  - Meshed AS topologies possible
  - Used today – discussed in the following sections!!!
  - RFC 1771

 IP Routing, v6.2 268

BGP is a distance vector protocol. This means that it will announce to its neighbors those IP networks that it can reach itself. The receivers of that information will say “if that AS can reach those networks, then I can reach them via it”.

If two different paths are available to reach one and the same IP subnet, then the shortest path is used. This requires a means of measuring the distance, a metric. All distance vector protocols have such means. BGP is doing this in a very sophisticated way by using attributes attached to the reachable IP subnet.

BGP sends routing updates to its neighbors by using a reliable transport. This means that the sender of the information always knows that the receiver has actually received it. So there is no need for periodical updates or routing information refreshments. Only information that has changed is transmitted.

The reliable information exchange, combined with the batching of routing updates also performed by BGP, allows BGP to scale to Internet-sized networks.

BGP-4 at a Glance

- **Carried within TCP**
  - Manually configured neighbor-routers
  - Therefore reliable transport (port 179)
- **Neighbor routers establish link-state**
  - Hello protocol (60 sec interval)
- **Incremental Updates upon topology changes**
  - New routes are updated
  - Lost routes are withdrawn
- **Each route is assigned a policy and an AS-Path leading to that network**
  - Using attributes

 IP Routing, v6.2 269

A router which has received reachability information from a BGP peer, must be sure that the peer router is still there. Otherwise traffic could be routed towards a next-hop router that is no longer available, causing the IP packets to be lost in a black hole.

TCP does not provide the service to signal that the TCP peer is lost, unless some application data is actually transmitted between the peers. In an idle state, where there is no need for BGP to update its peer, the peer could be gone without TCP detecting it.

Therefore, BGP takes care of detecting its neighbors presence by periodically sending small BGP keepalive packets to them. These packets are considered application data by TCP and must therefore be transmitted reliably. The peer router must also, according to the BGP specification, reply with a BGP keepalive packet.

Path Vector Protocol

- **Metric: Number of AS-Hops**
- **All traversed ASs are carried in the AS-Path attribute**
  - BGP is a "Path Vector protocol"
  - Better than Distance Vector because of inherent topology information
  - No loops or count to infinity possible

Each BGP update consists of one or more IP subnets and a set of attributes attached to them. The intrinsic metric is the number of AS hops. Note that this metric is given implicitly by a AS path attribute, which is a vector of all ASs traversed.

BGP Database

- **BGP routers also maintain a BGP Database**
  - Roadmap information through path vectors
  - Attributes
- **Routing Table calculated from BGP Database**
- **CPU/Memory resources needed**

The designers of the BGP protocol have succeeded in creating a highly scalable routing protocol, which can forward reachability information between Autonomous Systems, also known as Routing Domains. They had to consider an environment with an enormous amount of reachable networks and complex routing policies driven by commercial rather than technical considerations.

TCP, a well-known and widely proven protocol, was chosen as the transport mechanism. That decision kept the BGP protocol simple, but it put an extra load on the CPU or the routers running BGP. The point-to-point nature of TCP might also introduce a slight increase in network traffic, as any update that should be sent to many receivers has to be multiplied into several copies, which are then transmitted on individual TCP sessions to the receivers.

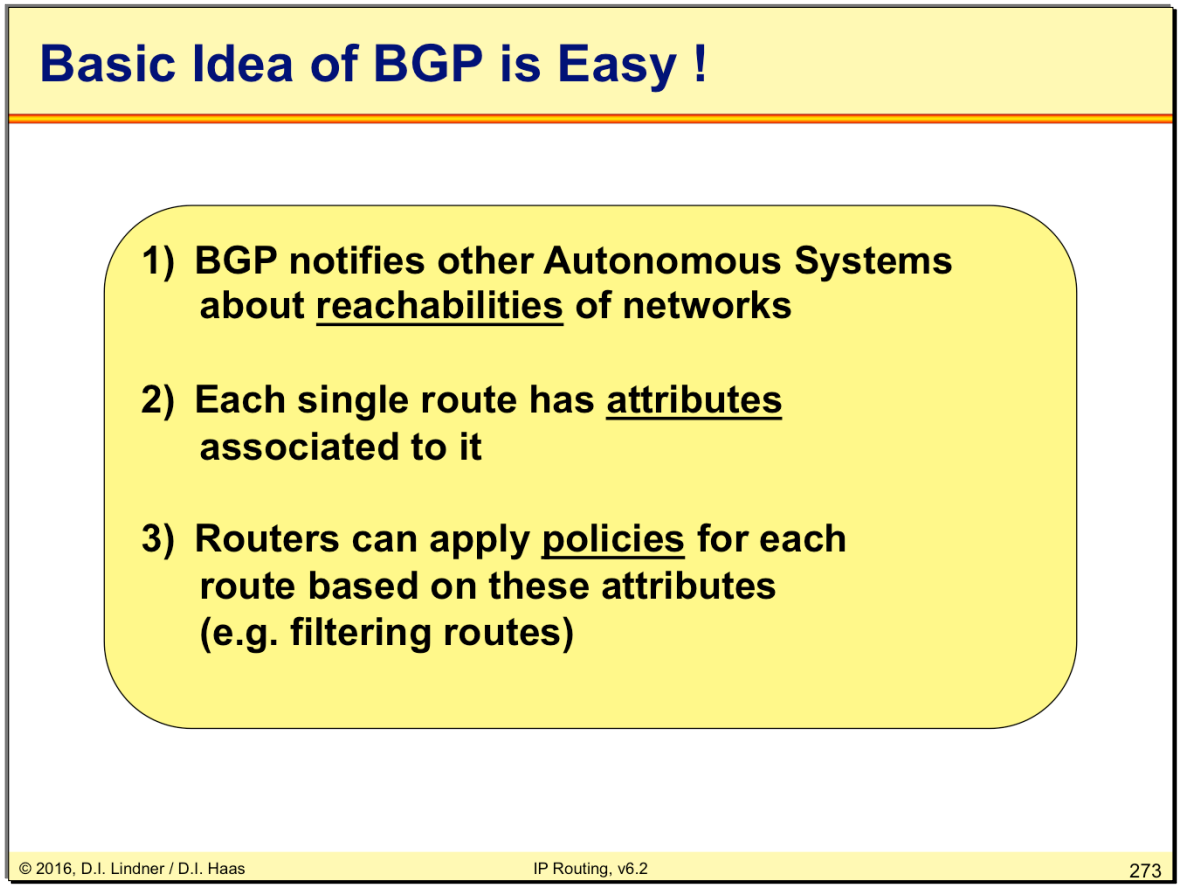
Whenever there was a design choice between fast convergence and scalability, scalability was the top priority. Batching of updates and the relative low frequency of keepalive packets are examples where convergence time has been second to scalability.

Some Interesting Numbers

- **Today's Internet BGP Backbone Routers are burdened**
  - About 415,000 routes (May 2012)
  - About 10,000 Autonomous Systems
- **Although excessive CIDR, NAT, and Default Routes**
- **Collapse expected**
  - Looking for new solutions

Internet routers do a hard job. The number of networks is increasing exponentially since the early 1990s and the only way to overcome routing table exhaustion is to apply excessive supernetting (CIDR), NAT, and default routing. In 2001 about 100,000 routes have been counted in typical BGP Internet router. Moreover, 10,000 ASs have been registrated.

Although this techniques significantly reduce the table growths a collapse is expected to happen in the near future—unless other techniques will be explored.



The text above summarizes the basic BGP-4 functionality. As it can be seen its not so complicated as many people think.

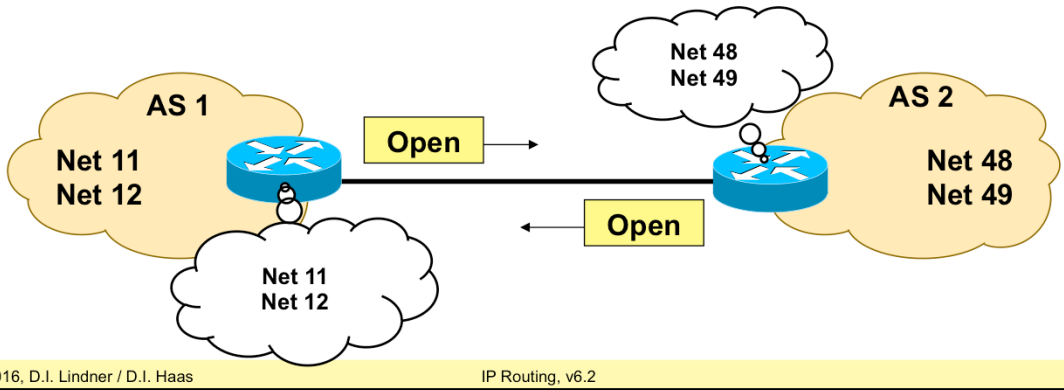
BGP Limitations

- **Destination based routing**
  - No policies for source address
- **Hop-by-hop routing**
  - Leads to hop-by-hop policies
  - Connectionless nature of IP
  - Mitigated through
    - Community attribute
    - Peer groups

There are still some limitations in BGP. It is impossible to implement source address-based policies with BGP (unless supported by vendor specific techniques). Furthermore BGP is still hop-by-hop routing, that is, the connectionless nature of IP makes it impossible to foresee what the next routers will do with the route.

Neighborship Establishment

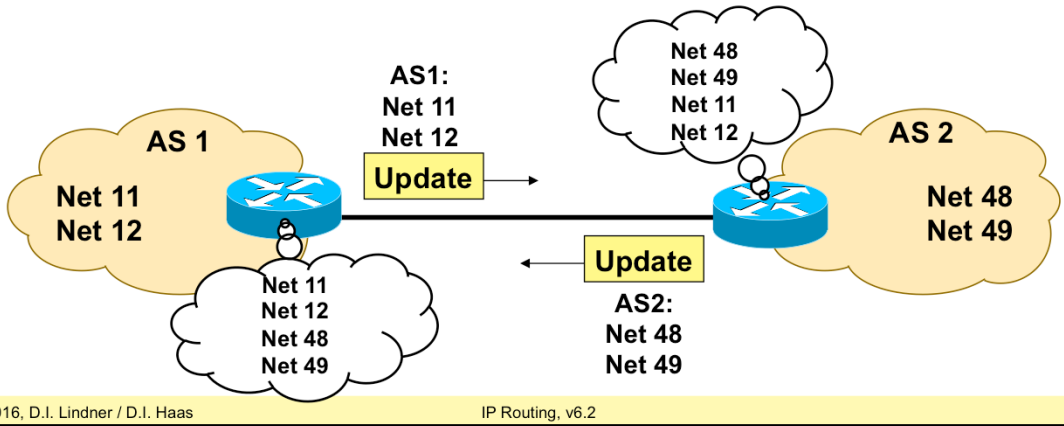
- **Open Message**
  - BGP Version (4)
  - AS number
  - BGP Router-ID (IP address)
  - Hold Time
- **Problems are indicated with Notification message**



The BGP protocol is carried in a TCP session, which must be opened from one router to the other. In order to do so, the router attempting to open the session must be configured to know to which IP address to direct its attempts.

NLRI Update

- After open message, all known routes are exchanged using update messages
- Contains network layer reachability information (NLRI)
  - List of prefix and length

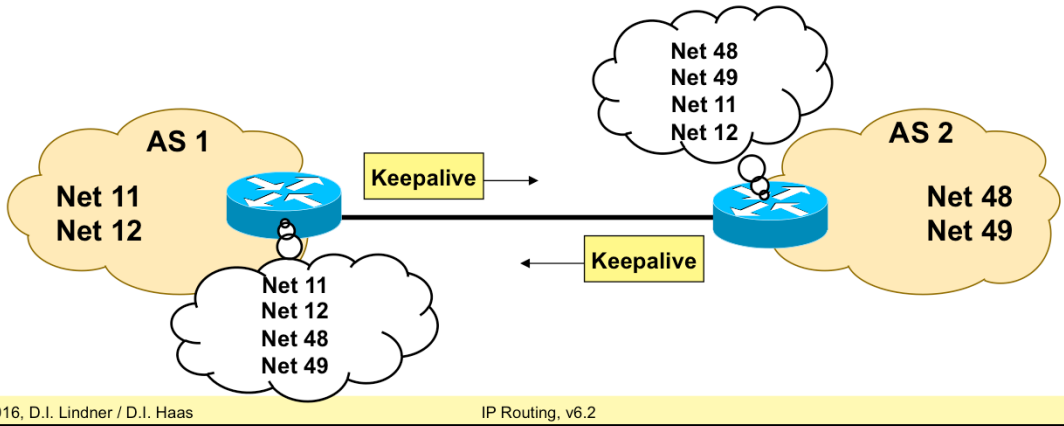


 IP Routing, v6.2 276

Once the BGP session is established, routing updates start to arrive. Each BGP routing update consists of one or more entries (routes). Each route is described by the IP address and subnet mask along with any number of attributes. The next-hop, AS-path and origin attributes must always be present. Other BGP attributes are optionally present.

Steady State

- After Open/Update procedure, BGP is nearly quiet – *No periodic updates* !
- Only keepalive messages are sent
  - 19 Bytes
  - Per default every 60s

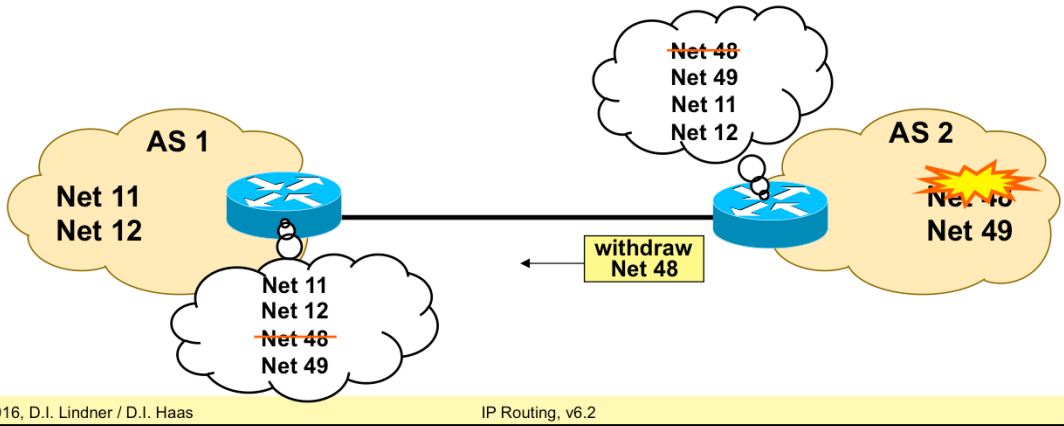


 IP Routing, v6.2 277

After finishing the update process, no periodic updates are sent, just keepalives by default every 60 seconds

Topology Change:

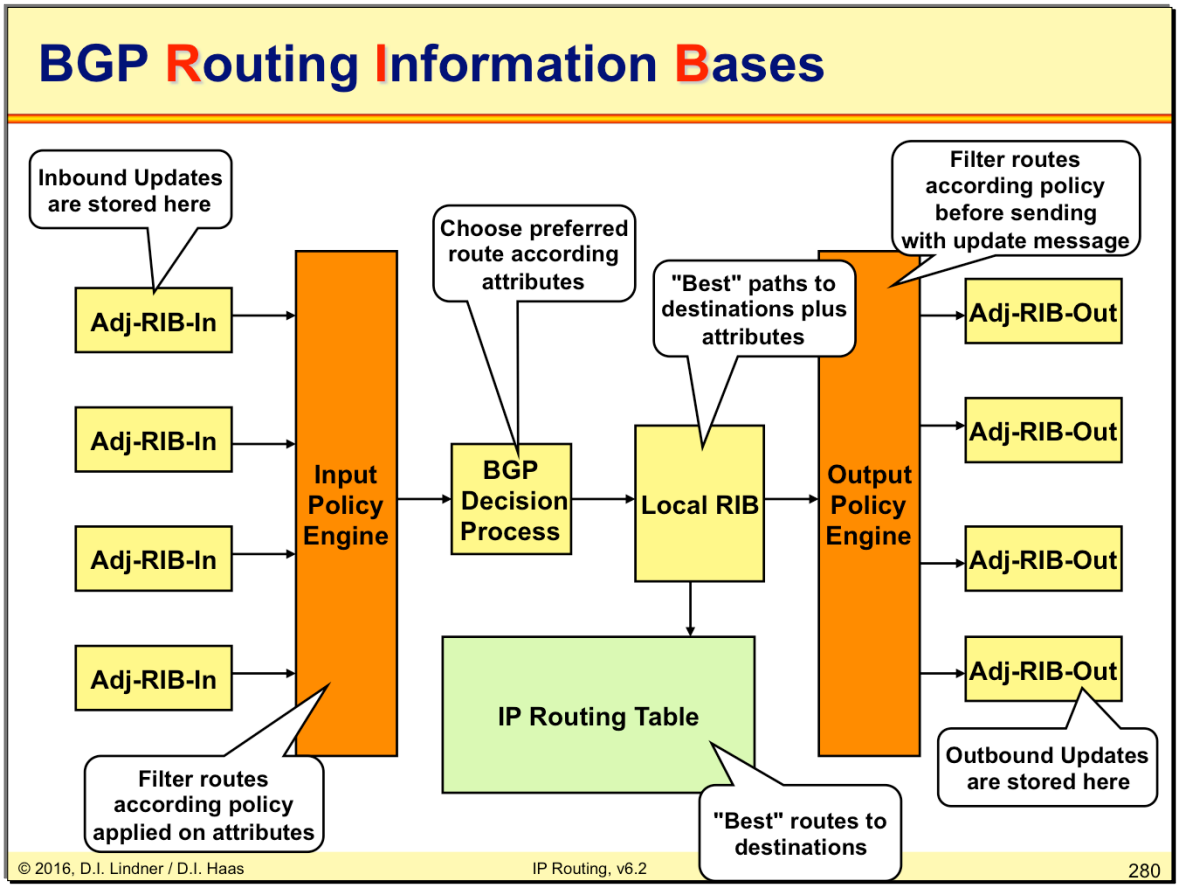
- Incremental Updates upon topology or attribute changes
- Withdraw message upon loss of network



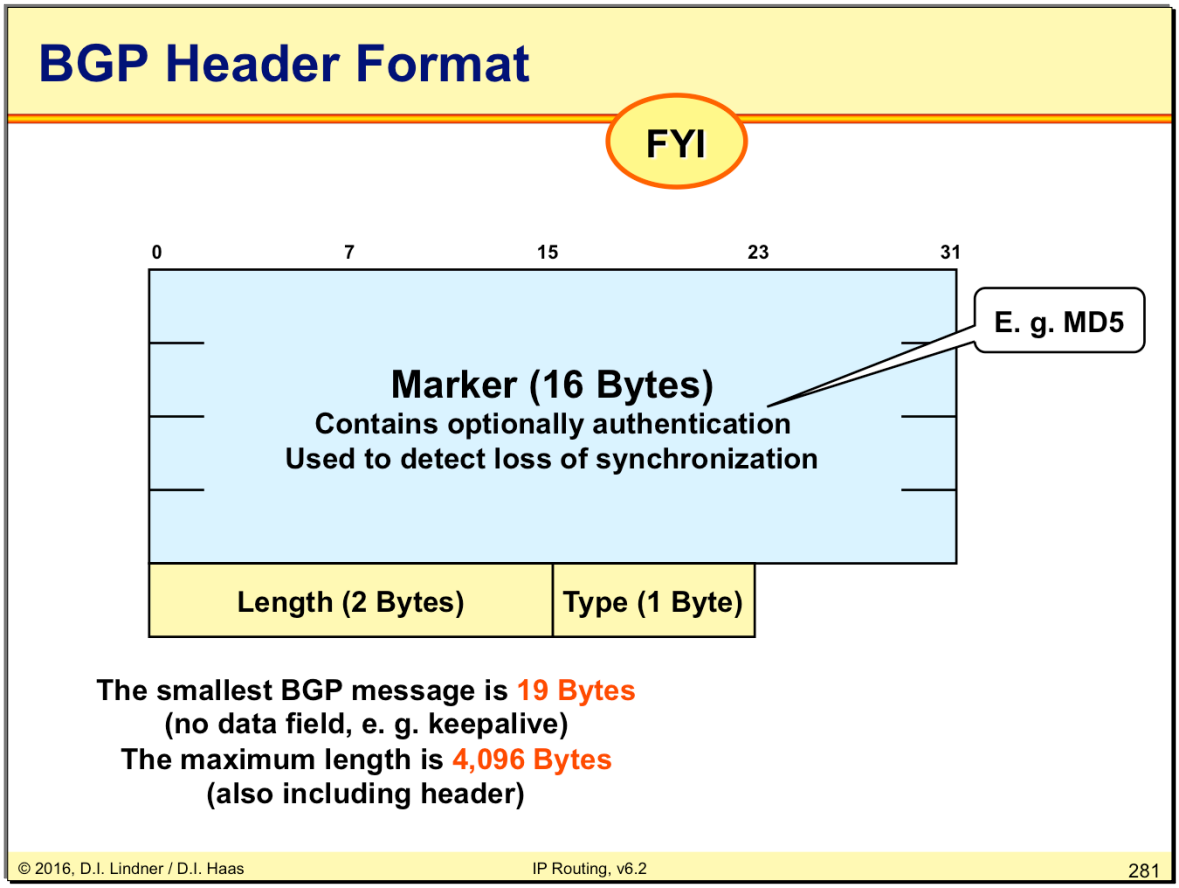
If there is a topology change, only information about the changes is transmited.

RIB

- **BGP routing information is stored in RIBs**
- **RIBs might be combined (vendor specific)**
- **Only best paths are forwarded to the neighboring ASs**
- **Alternative paths remain in the BGP table**
  - "Feasible routes" in Adj-RIB-In
  - Are used if the original path is withdrawn



The Adj-RIB-In maintains also feasible routes, whereas only the best route is kept in the Local RIB. In case of a withdrawn message for this single best route, the best feasible route becomes active.



This is the basic BGP header format. This and the following slides marked with a "FYI" at the upper-left of the slide are only given "for your information". It is usually not necessary to know this details by heart—unless you plan to go deeper in BGP.

Message types:

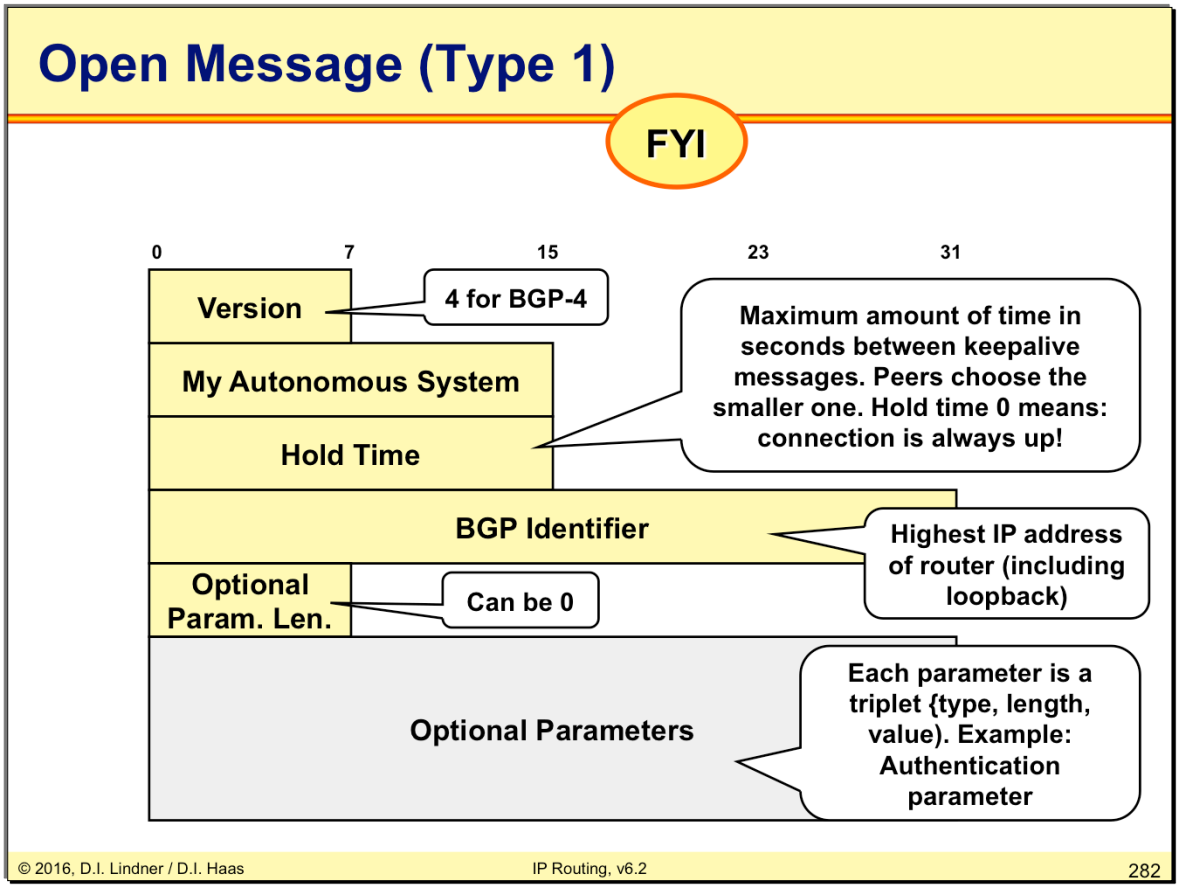
Open (type 1) to establish relationship between BGP neighbors

Update (type 2) to advertise reachability information with its corresponding path attributes. Path attributes are used for BGP route decision process and supports establishing of routing policy between ASs.

Notification (type 3) to report errors to the neighbor. After notification is sent relationship will be terminated.

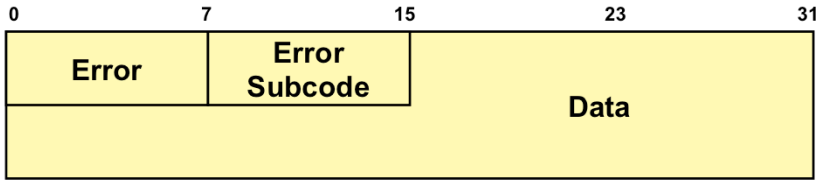
Keepalive (type 4) to constantly monitor reachability of BGP neighbor.

Route Refresh (type 5, RFC 2918) to enforce a re-advertisement from the Adj-RIB-out from a BGP neighbor. Adj-RIB-out = storage place for all BGP-routes already sent to BGP neighbors.



Notification Message (Type 3)

FYI



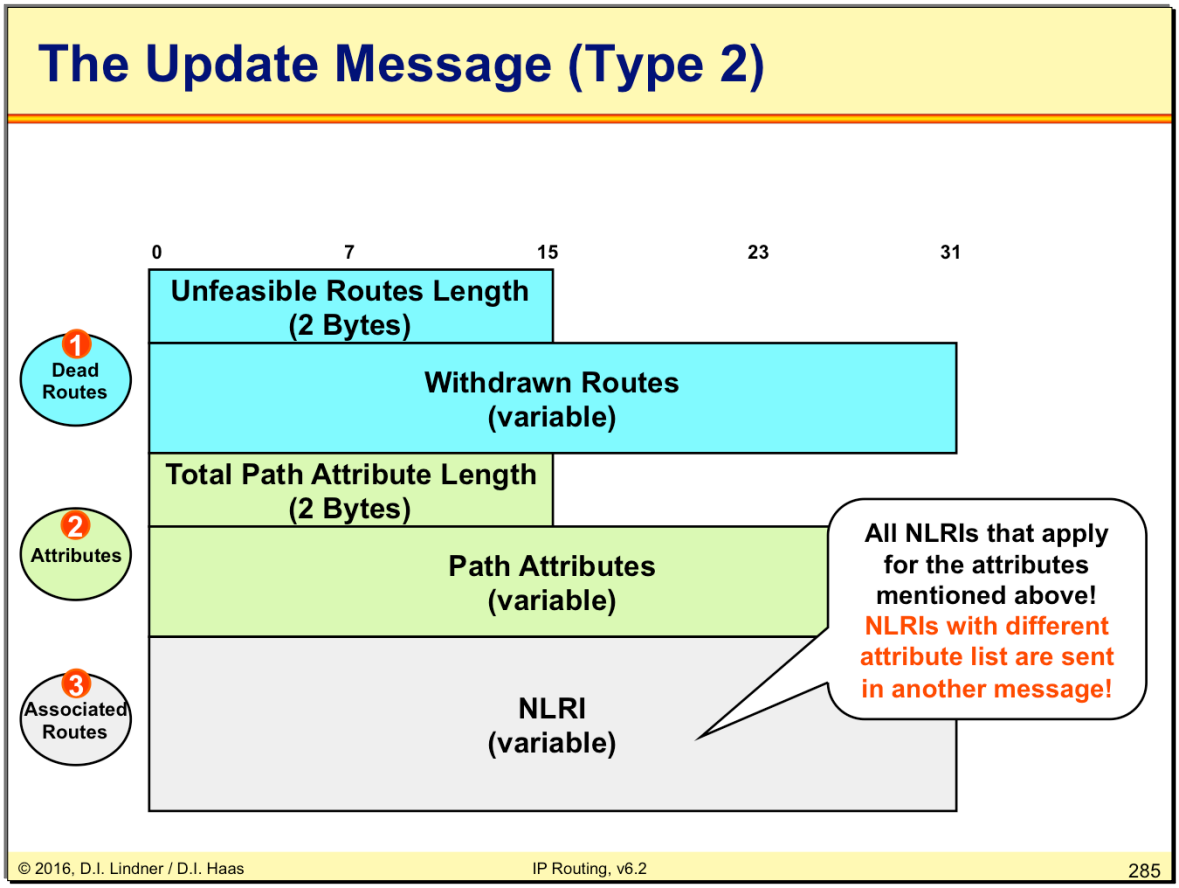
Notification is always sent when an error is detected.

After that, the connection is closed.

Keepalive Message (Type 4)

- Consists of header only (19 bytes)
- Must be sent before hold time expires
- Recommended keepalive rate
  = 1/3 of hold time
- Not necessary if update message is sent

Keepalive messages are sent periodically, by default at 60 seconds interval.

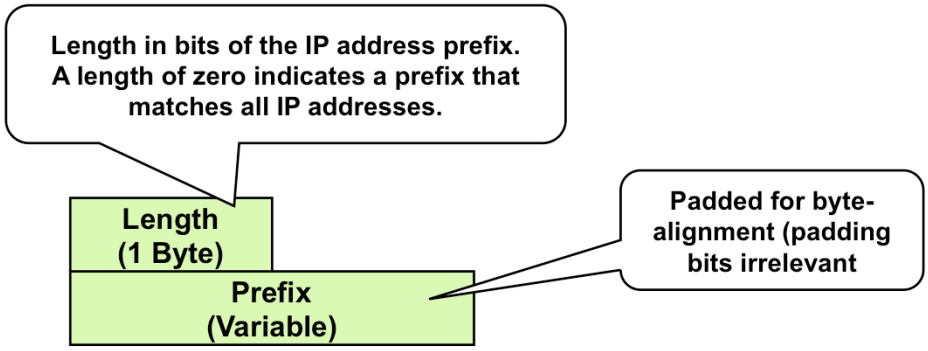


The picture above shows the most important message within BGP: the update. Note that the update message consists of three parts. The first part contains all unfeasible routes (also known as "withdrawn" routes). The second part contains a consistent set of attributes for the following regular routes listed in the third part of the message.

Note that another update message has to be sent if a route (NLRI) should be advertised with a different set of attributes.

Withdrawn Routes

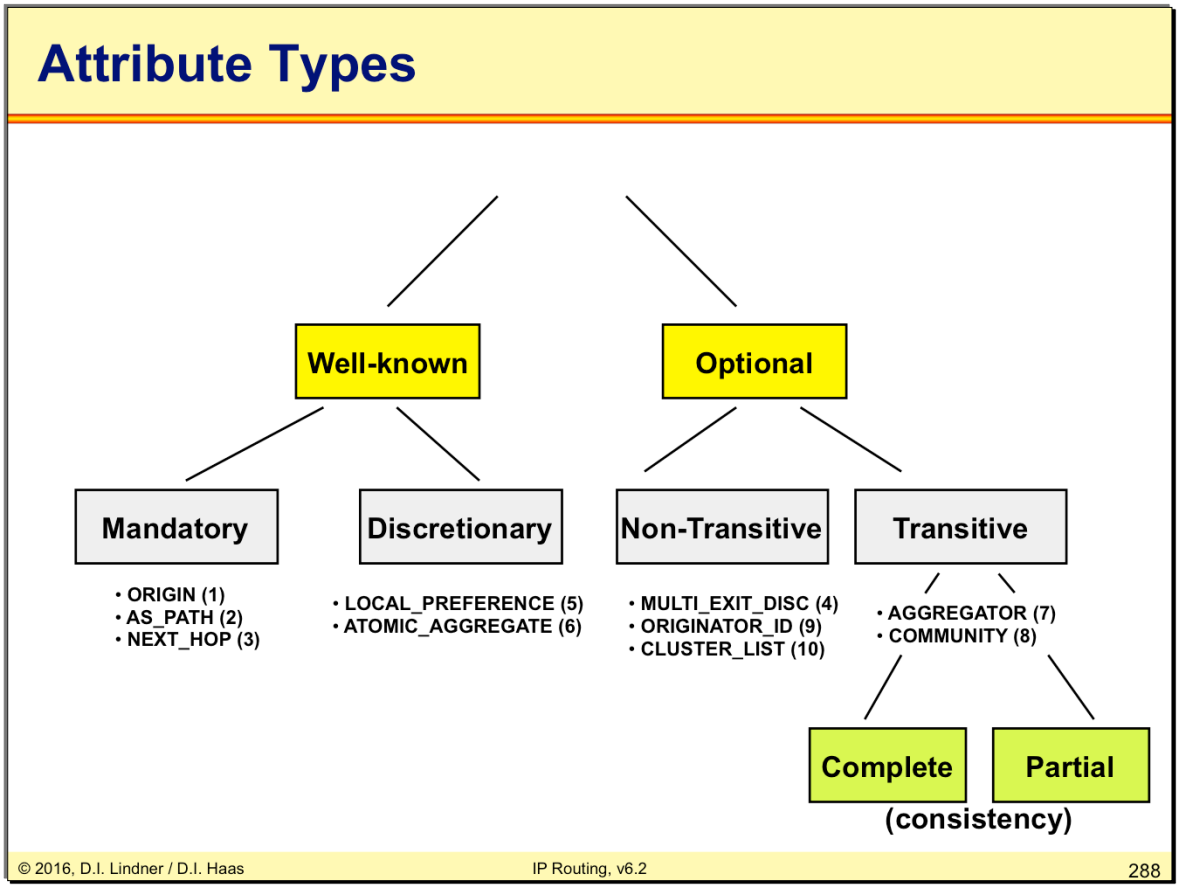
FYI



...How destinations are specified within an update

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
  - Introduction
  - BGP Basics
  - BGP Attributes
  - BGP Special Topics
  - CIDR



Each BGP update consists of one or more IP subnets and a set of attributes attached to them. Some of the attributes are required to be recognized by all BGP implementations. Those attributes are called well-known BGP attributes.

Attributes that are not well known are called optional. These could be attributes specified in a later extension of the BGP protocol or even private vendor extensions not documented in a standard document.

Well-known must be recognized by all BGP implementations.

Well-known mandatory must be included in every Update message (e.g. Origin, AS_Path, Next_Hop).

Well-known discretionary may or may not be included in every Update message (e.g. Local_Preference, Atomic_Aggregate).

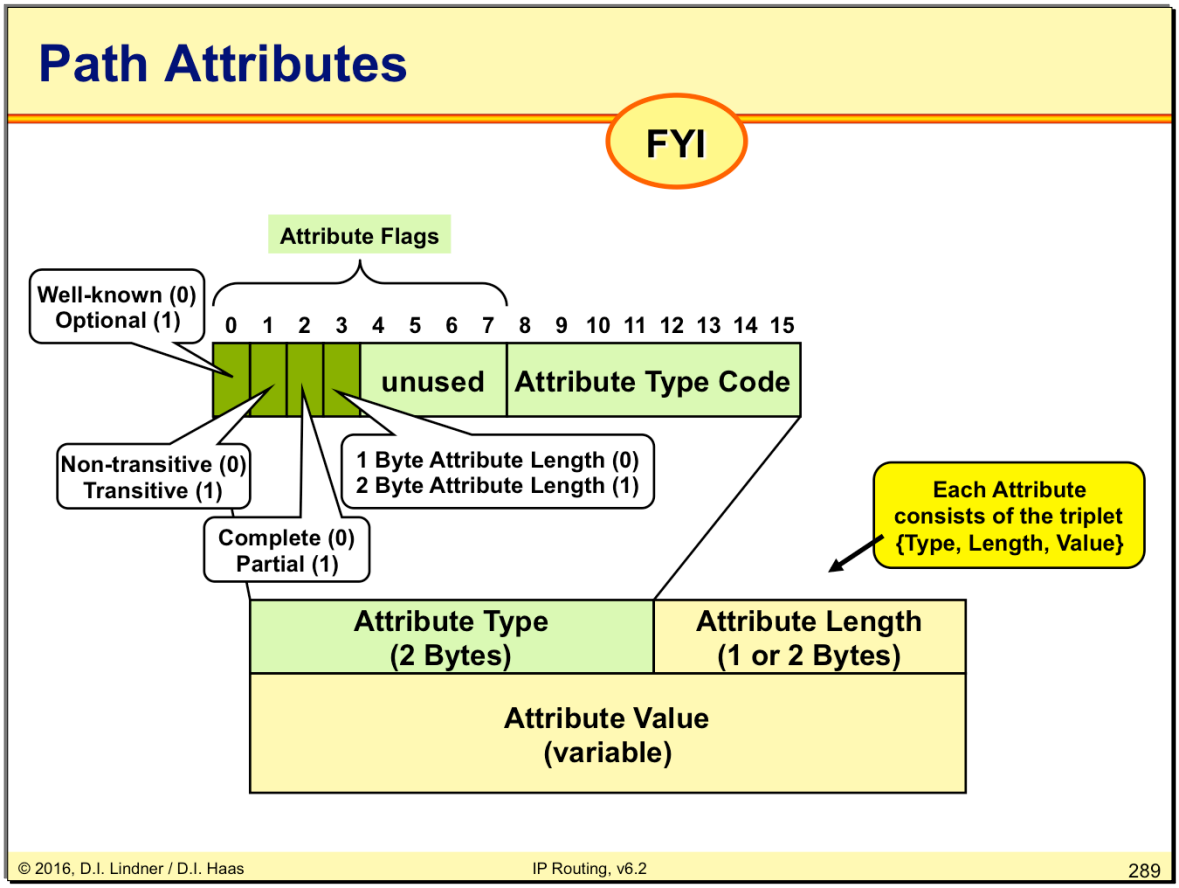
All well-known attributes must be passed along to other BGP peers. Some will be updated properly first, if necessary.

Optional: it is not required or expected that all BGP implementation support all optional attributes. May be added by the originator or any AS along the path. Paths are accepted regardless whether the BGP peer understands an optional attribute or not.

Handling of recognized optional attributes: Propagation of attribute depends on meaning of the attribute. Propagation of attribute is not constrained by transitive bit of attribute flags but depends on the meaning of the attribute.

Handling of unrecognized optional attribute: Propagation of attribute depends on transitive bit of attribute flags. If transitive bit is set paths are accepted (attribute is ignored) and attribute remains unchanged when path is passed along to other peers. Attribute is marked as partial (bit 3 of attribute flags). Example: Community

If transitive bit is not set hence non-transitive: Paths are accepted, attribute is quietly ignored and discarded when path is passed along to other peers. Example: Multi_Exit_Discriminator



Each attribute consists of a so called TLVs – Type, Length, Value.

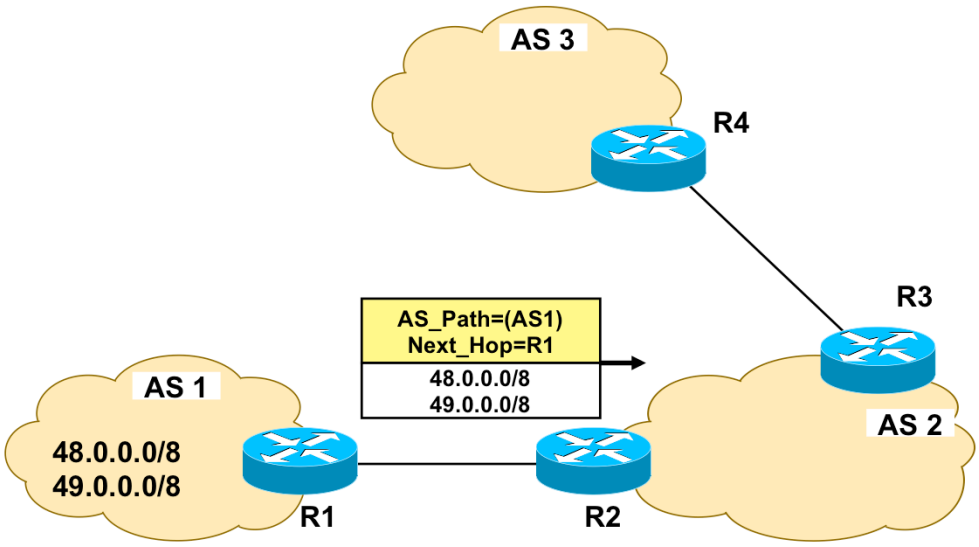
Well-known Mandatory

- **AS_Path contains all ASs traversed for this route**
- **Next_Hop indicates the last EBGP router leading to this route**
  - Not necessarily the physical next hop
- **Origin indicates how this route was learned**

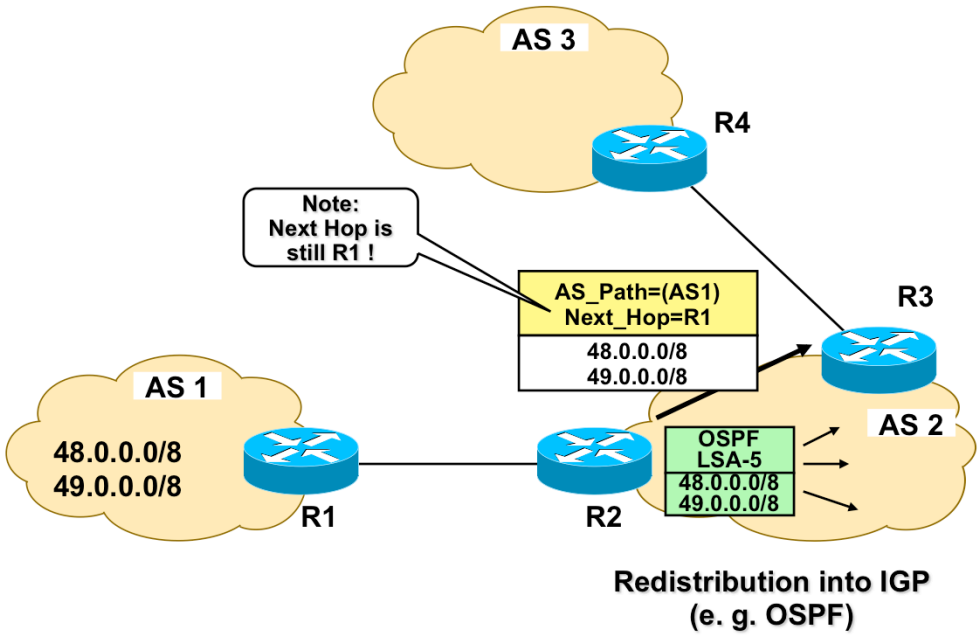
There is a small set of three specific well-known attributes that are required to be present on every update. These three are the AS-path, next-hop and origin attributes. They are referred to as well-known mandatory attributes.

Other well-known attributes may or may not be present depending on the circumstances under which the updates are sent and the desired routing policy. The well-known attributes that could be present, but are not required to be present, are called well-known discretionary attributes.

Path Vector Protocol (1)

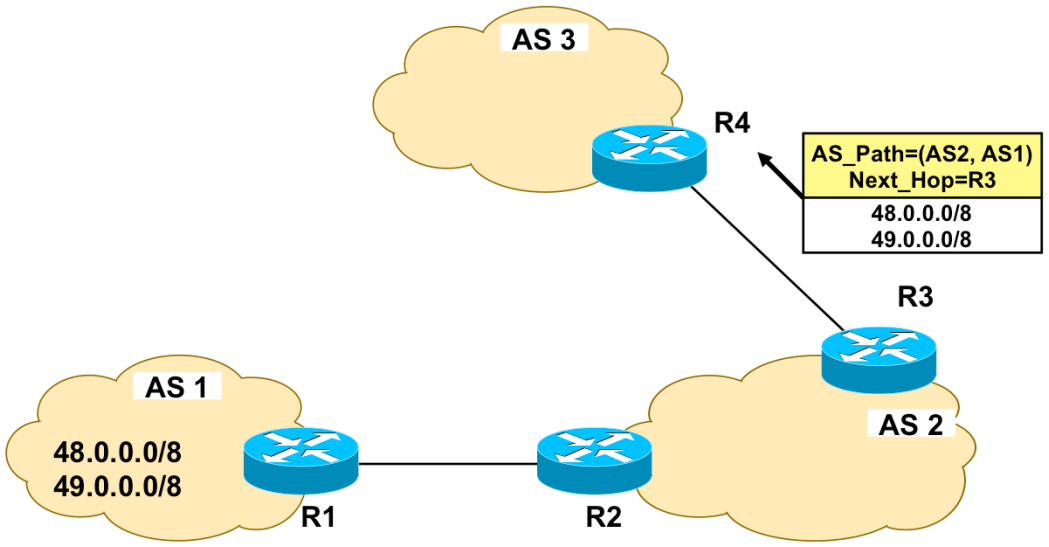


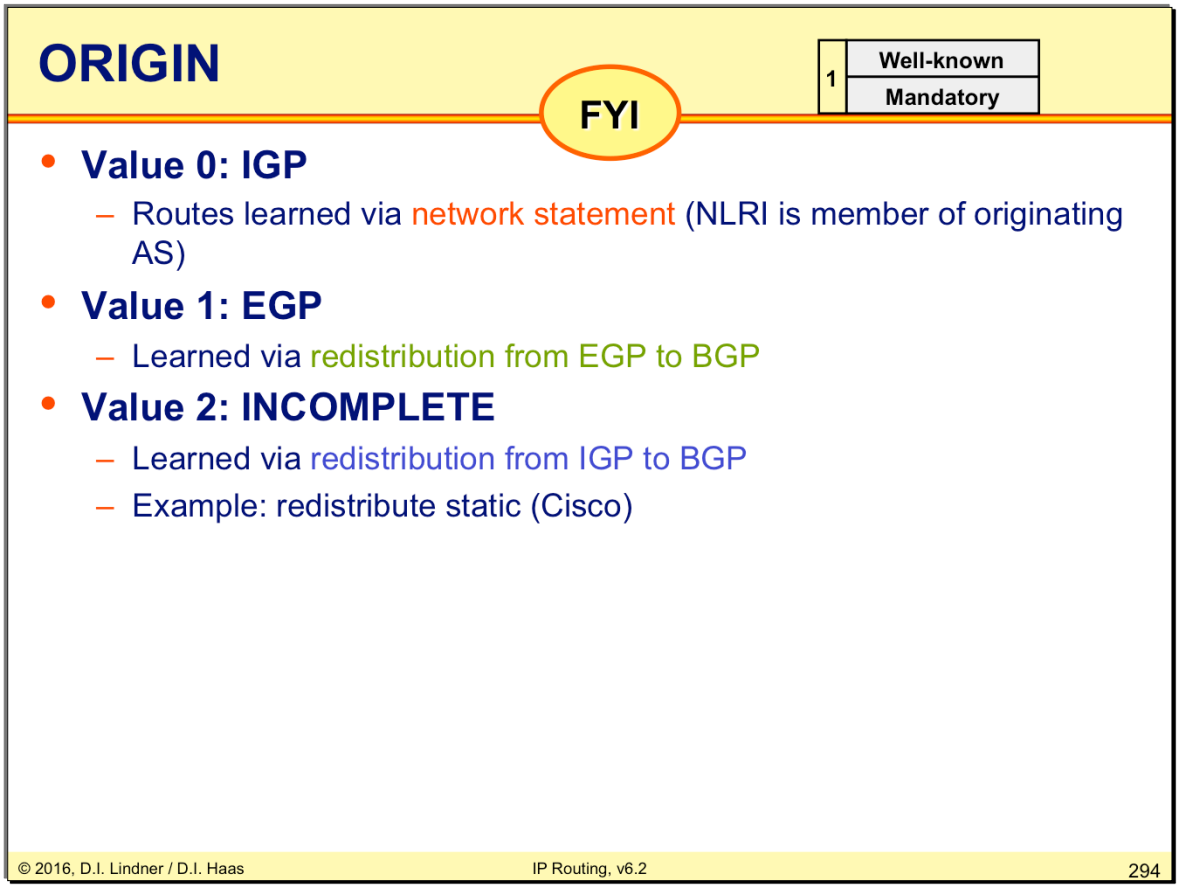
Path Vector Protocol (2)



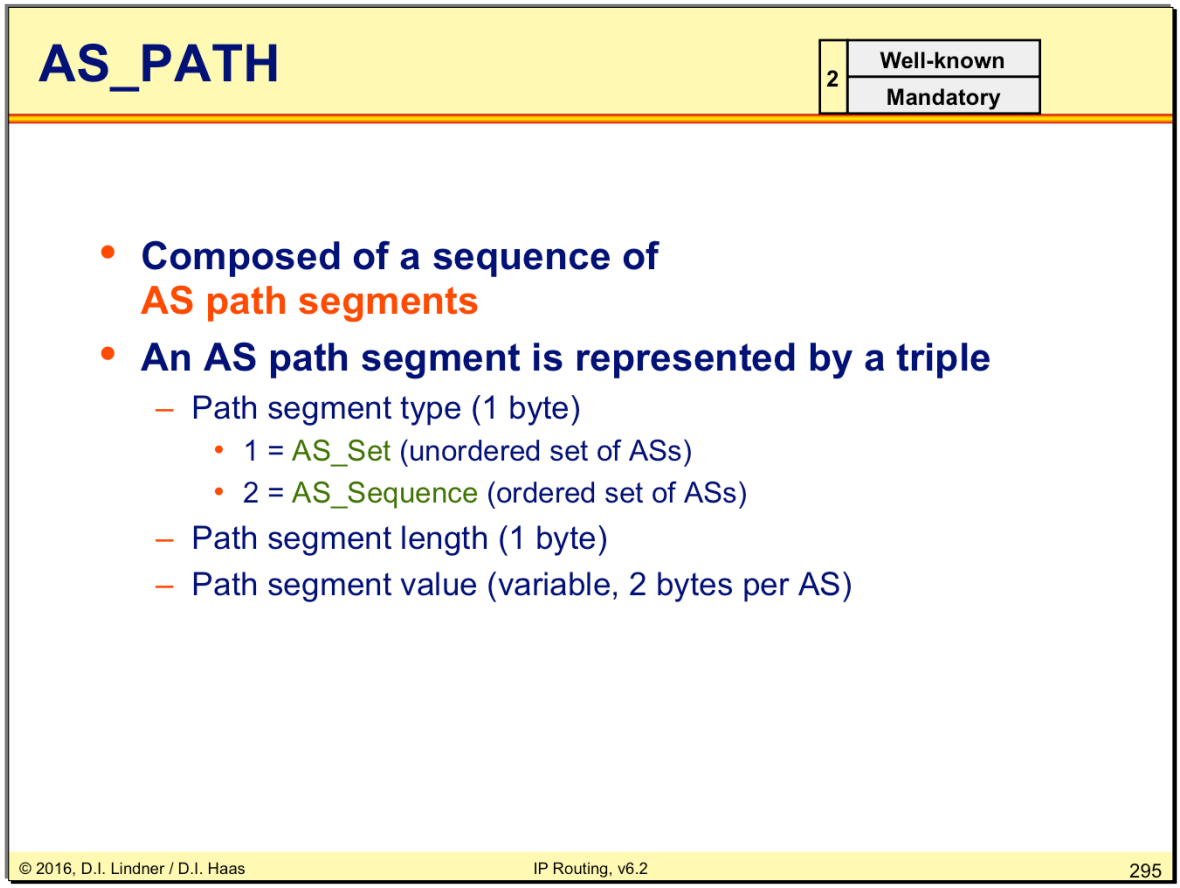
 IP Routing, v6.2 292

Path Vector Protocol (3)





The origin attribute is set when the route is first injected into the BGP. If information about an IP subnet is injected using the network command or via aggregation (route-summarization within BGP) the origin attribute is set to IGP. If the IP subnet is injected using redistribution, the origin attribute is set to unknown or incomplete (these two words have the same meaning). The origin code, EGP, was used when the Internet was migrating from EGP to BGP and is now obsolete.



The AS-path attribute is modified each time the information about a particular IP subnet passes over an AS border. When the route is first injected into the BGP the AS-path is empty.

Each time the route crosses an AS boundary the transmitting AS prepends its own AS number to appear first in the AS-path. The sequence of ASs, through which the route has passed, can therefore be tracked using the AS-path attribute.

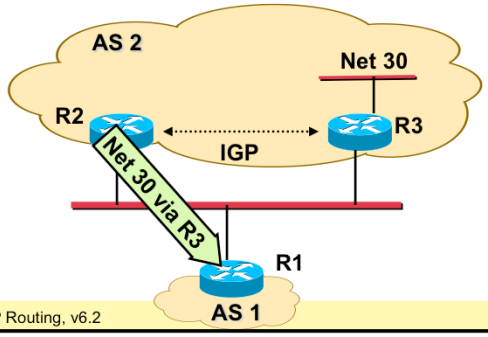
Who is NEXT_HOP?

3	Well-known
	Mandatory

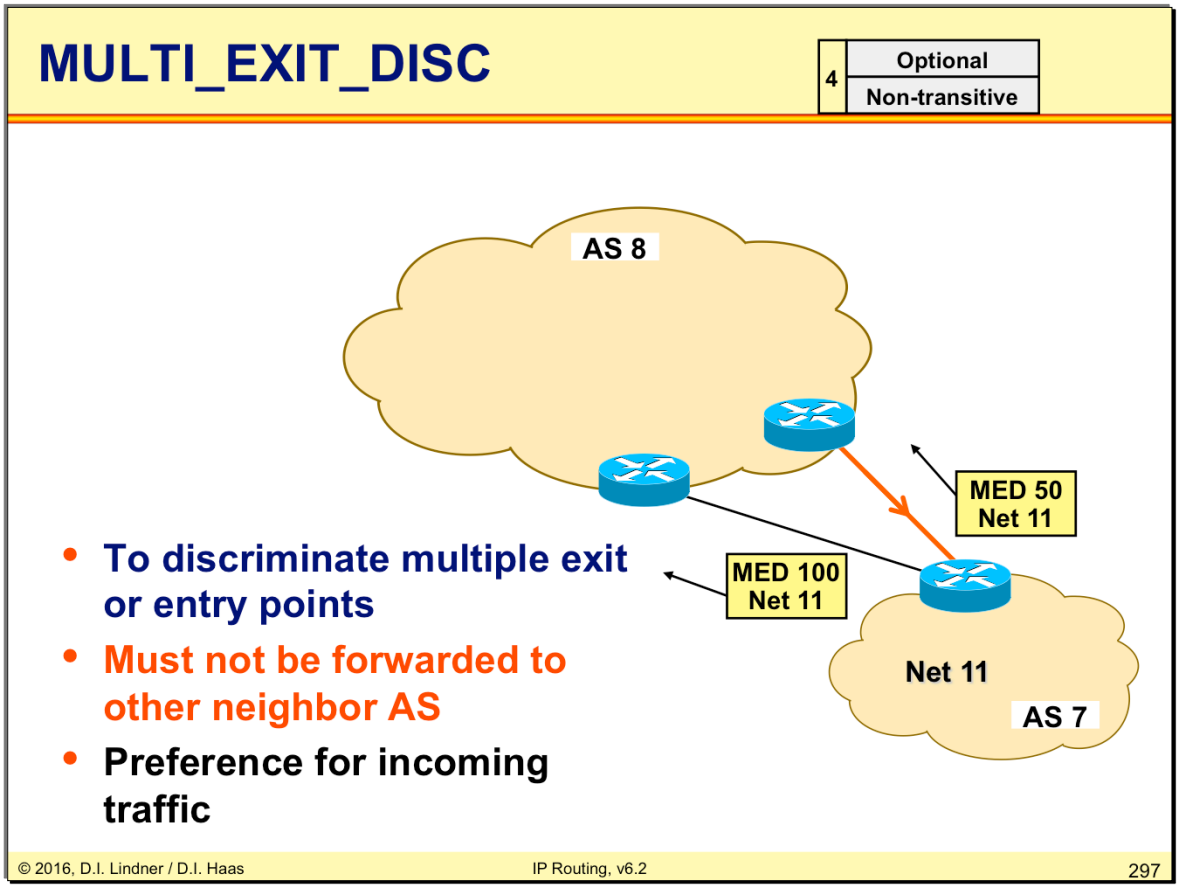
- **The boundary router that advertized the route in this AS is the next hop**
  - Recursive routing table lookup might be necessary to determine the true physical next hop
- **Exception:**
  - On multi-access media (Ethernet, FDDI) always the physical next hop must be indicated

R1 and R2 have BGP session established, R3 speaks IGP only.

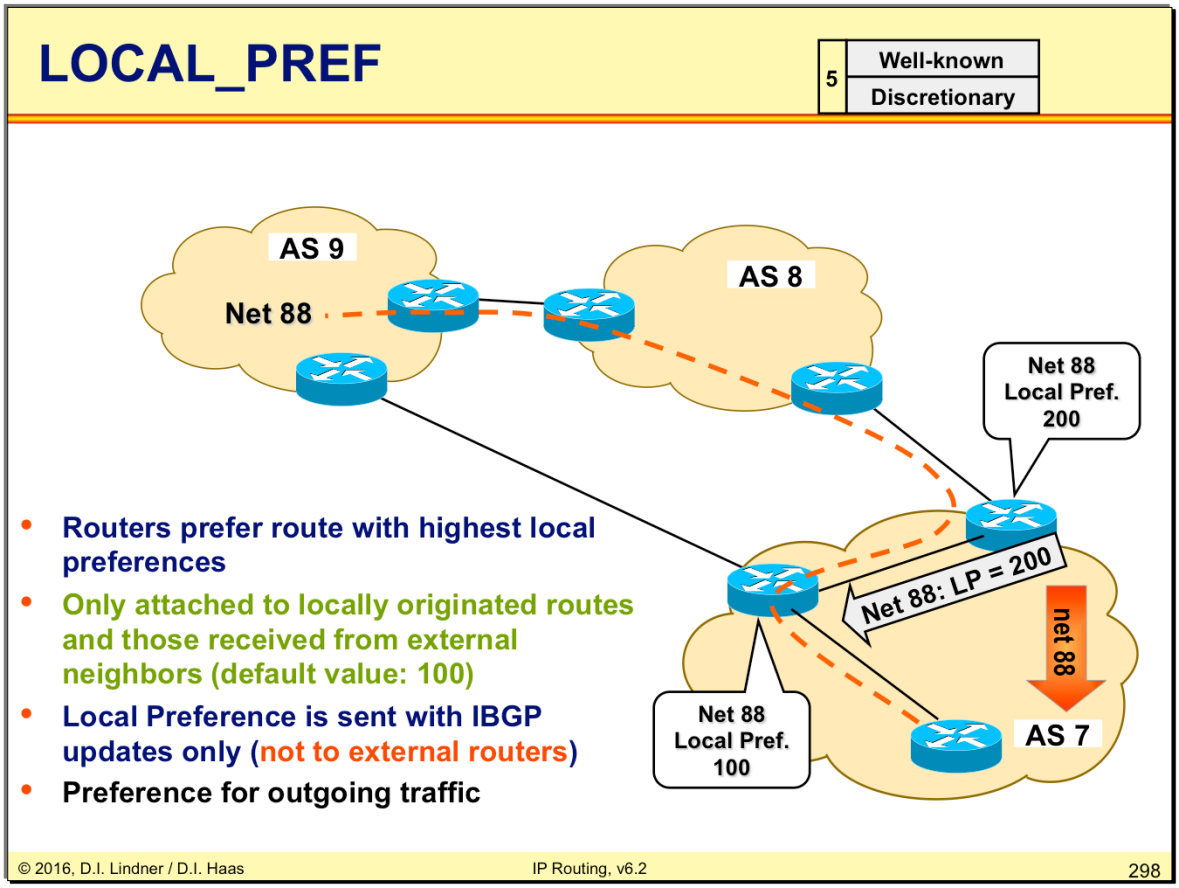
R2 advertises R3 as next hop to Net 30 because R3 is on the same physical media.



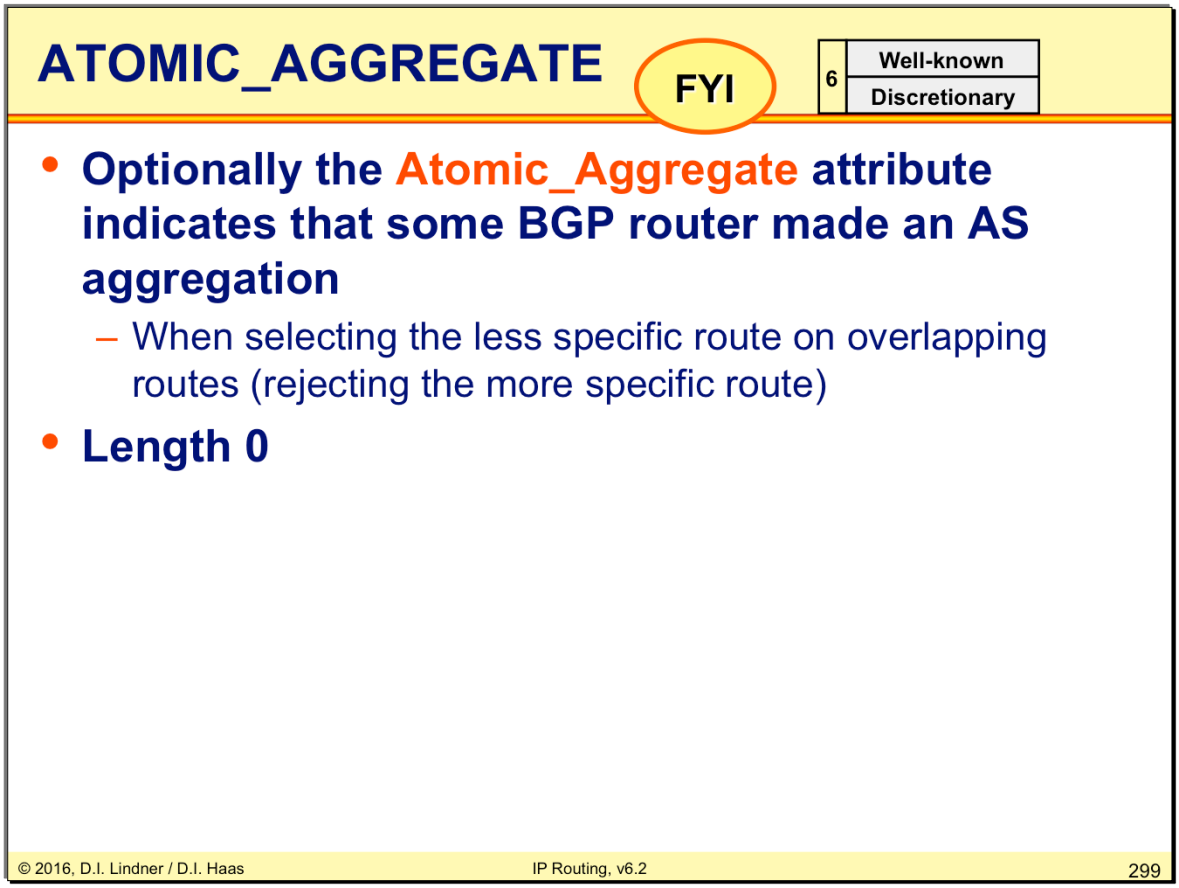
The next-hop attribute is also modified as the route passes through the network. It is used to indicate the IP address of the next-hop router—the router to which the receiving router should forward the IP packets toward the destination advertised in the routing update.



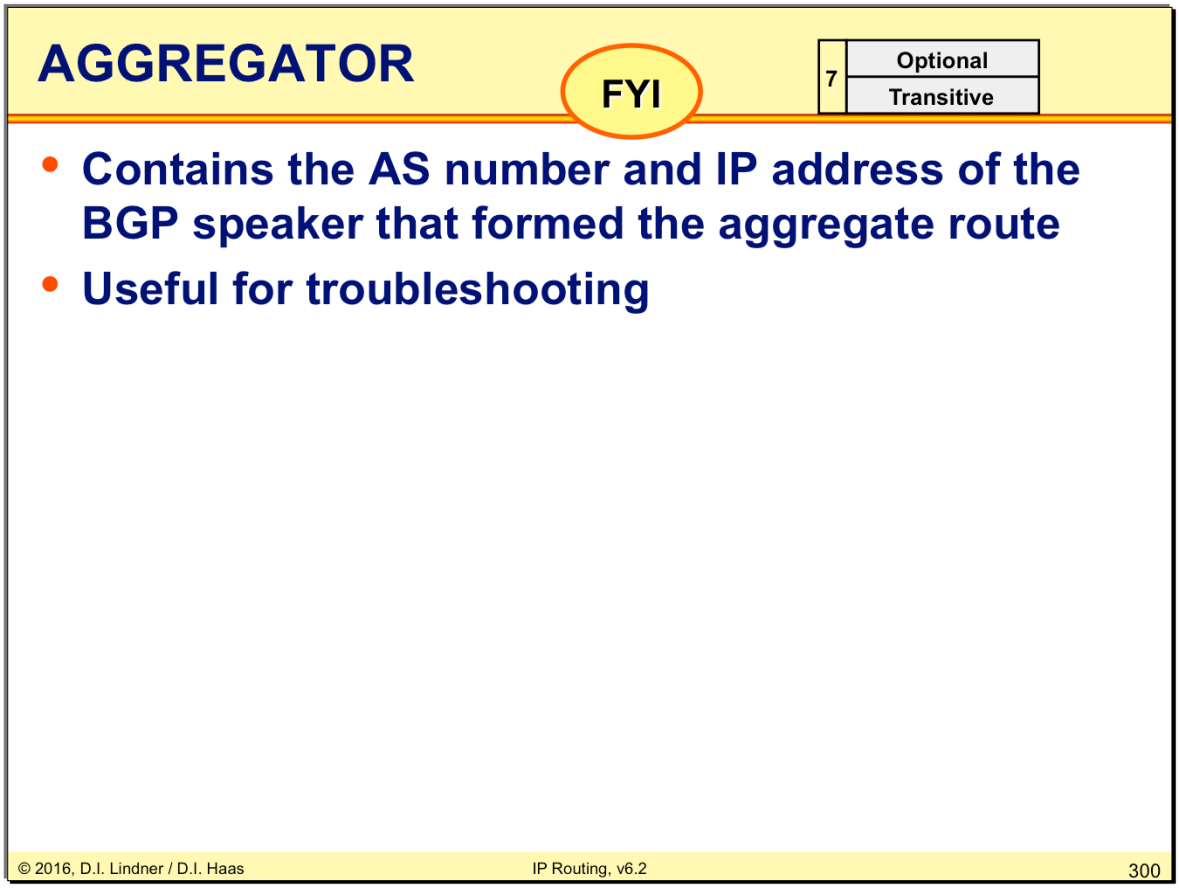
One of the non-transitive optional attributes is the Multi-Exit-Discriminator (MED) attribute which is also used in the route selection process. Whenever there are several links between two adjacent ASs, multi-exit-discriminator may be used by one AS to tell the other AS to prefer one of the links over the other for specific destinations.



Local Preference is used in the route selection process. The attribute is carried within an AS only. A route with a high local preference is preferred over a route with a low value. By default, routes received from peer AS are tagged with the local preference set to the value 100 before they are entered into the local AS. If this value is changed through BGP configuration, the BGP selection process is influenced. Since all routers within the AS get the attribute along with the route, a consistent routing decision is made throughout the AS. In our example the right router will insert net 88 into the local IGP in order to get traffic for net 88. The left router will not attract traffic for net 88 even if the AS Path is shorter. The left router will only take over, if the right router stops or is disconnected from AS8.



The Atomic Aggregate attribute is attached to a route that is created as a result of route summarization (called aggregation in BGP). It signals that information that was present in the original routing updates may have been lost when the updates where summarized into a single entry.



Aggregator identifies the AS and the router within that AS that created a route summarization, aggregate.

COMMUNITY

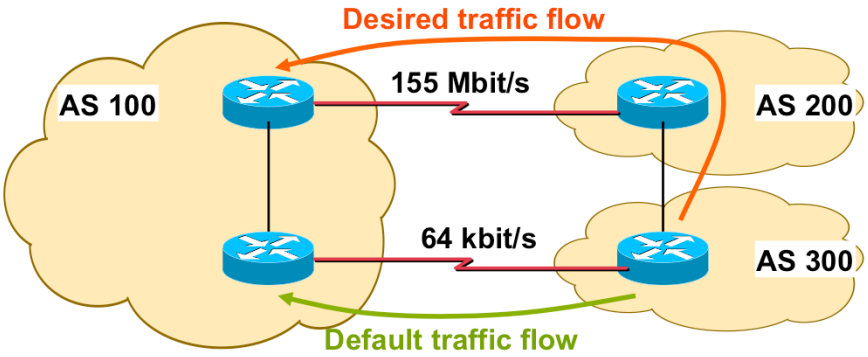
8	Optional
	Transitive

- **Group of destinations that share a common policy**
  - Each destination could be member of multiple communities
  - Carried across ASs
- **Community strings are simple policy labels**
  - Any BGP router can tag routes in incoming and outgoing routing updates or when doing redistribution
  - Any BGP router can filter routes in incoming or outgoing updates or select preferred routes based on communities

 IP Routing, v6.2 301

A Community is a numerical value that can be attached to certain routes as they pass a specific point in the network. The community value can then be checked at other points in the network for filtering or route selection purposes. BGP configuration may cause routes with a specific community value to be treated differently than others.

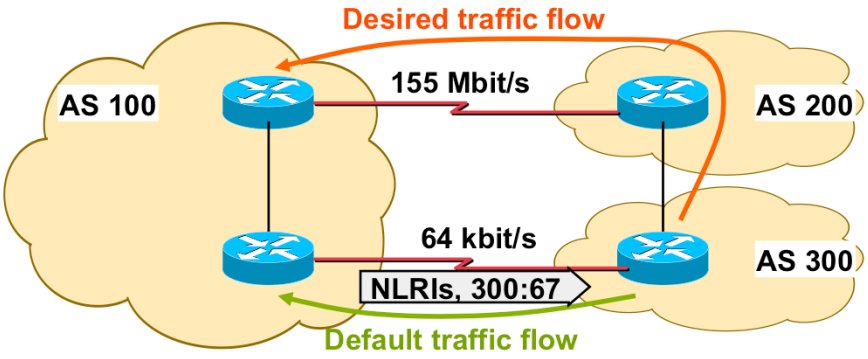
Community Example (1)



- **Assume AS 100 wants AS 300 to use the 155 Mbit/s link to reach own networks**
  - MED: not possible (non-transitive)
  - Local Preference: will admin of AS 300 set it?
- **Best and easiest: Use community !**

The picture above gives an example where the community could be implemented.

Community Example (2)



- Receiving a community string means "apply the predefined policy"
- In our example 300:67 means: "set local preference to 50"

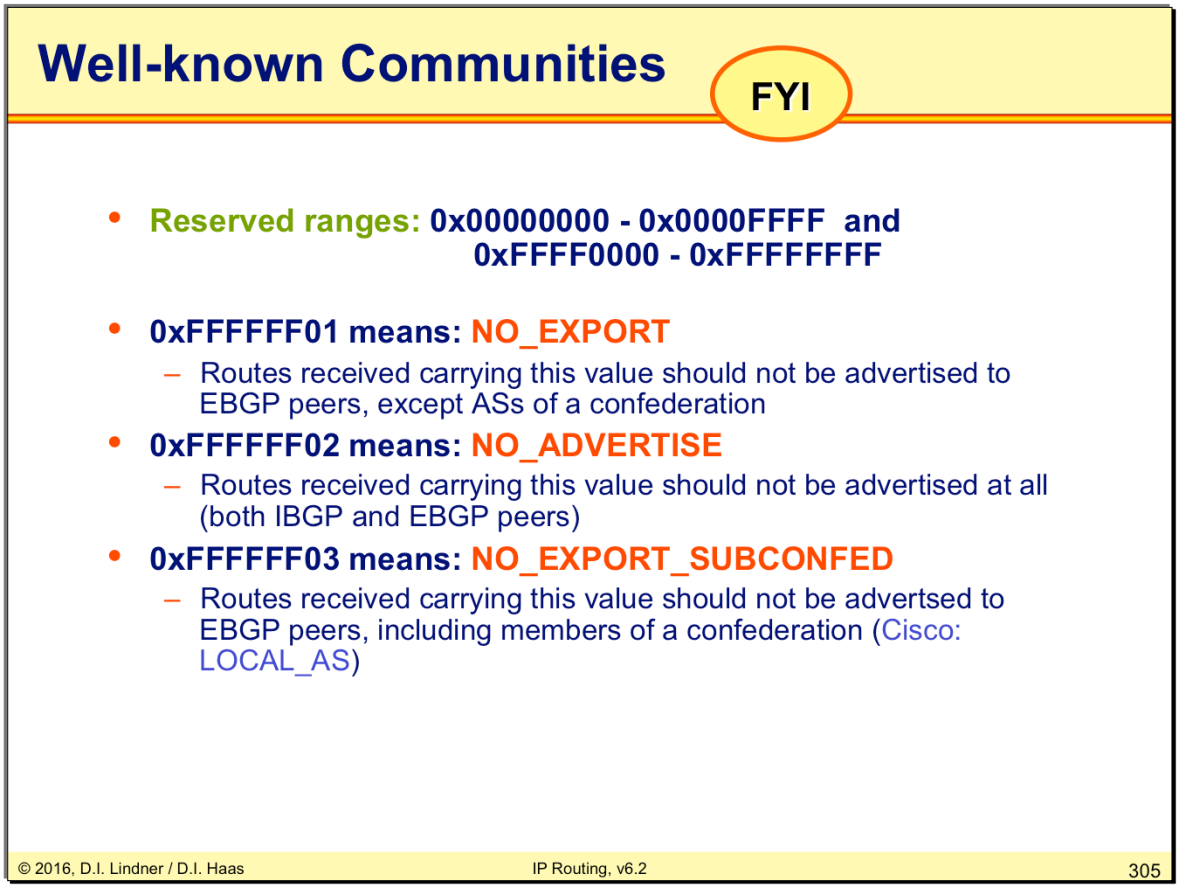
 IP Routing, v6.2 303

The picture above gives an example where the community could be implemented (continued from previous slide). By receiving the community 300:67 the local preference is set to 50. Remember: The default for local preference is 100. The higher the better. So if AS300 receives an update from AS200 about net 300 then this update has better local preference than the update from AS100.

Defining Communities

FYI

- **More than one BGP community per route allowed**
  - By default, communities are stripped in outgoing BGP updates
- **Private range:**
  0x00010000 - 0xFFFEFFFF
- **Common practice**
  - High order 16 bit: AS number
  - Low order 16 bit: Local significance



Easy to memorize: Values of all-zeroes and all-ones in high-order 16 bits are reserved.

Administrative Weight (Cisco)

FYI

- **No attribute – just a local parameter**
- **Applies only to routes within an individual router**
- **Number between 0 and 65535**
  - The higher the weight the more preferable the route
- **Initially invented to translate public routing policies (EGP)**

Note that the Administrative Weight is a Cisco specific attribute.

Decision Hierarchy

FYI

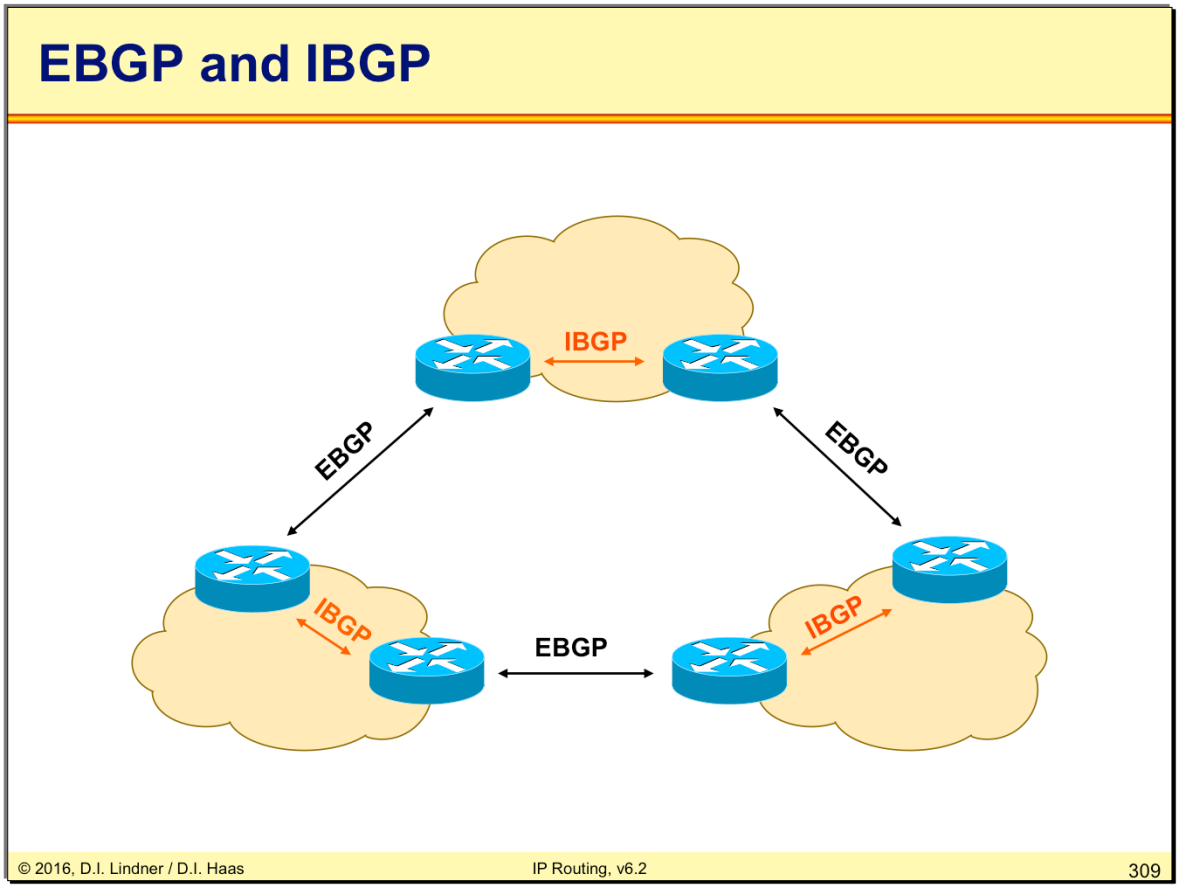
1. Prefer highest weight (Cisco)
2. Prefer highest local preference
3. Prefer locally originated routes
4. Prefer shortest AS-Path
5. Prefer lowest origin code
6. Prefer lowest MED
7. Prefer EBGP path over IBGP path
8. Lowest IGP metric to next hop
9. Prefer oldest route for EBGP paths
10. Prefer path with lowest neighbor BGP router ID

 IP Routing, v6.2 307

If routes have same local preference the route that was locally originated will be preferred. At last the BGP router ID can be used as tie-breaker.

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP)**
  - Introduction
  - BGP Basics
  - BGP Attributes
  - BGP Special Topics
  - CIDR



Interior BGP or "IBGP" allows edge routers to share NLRI and associated attributes, in order to enforce an AS-wide routing policy.

IBGP is responsible to assure connectivity to the "outside world" i. e. to other autonomous systems. That is, all packets entering this AS and were not blocked by policies should reach the proper exit BGP router. All transit routers inside the autonomous system should have a consistent view about the routing topology. Furthermore, IBGP routers must assure "synchronization" with the IGP, because packets cannot be continuously forwarded if the IGP routers have no idea about the route. Thus, IBGP routers must await the IGP convergence time inside the AS. Obviously this aspect assumes that BGP routes are injected to transit IGP routers by redistribution. The story with synchronization is explained a few slides later…

Internal and External BGP

- **EBGP messages are exchanged between peers of different ASs**
  - EBGP peers should be directly connected
- **Inside an AS this information is forwarded via IBGP to the next BGP router**
  - IBGP messages have same structure like EBGP messages
- **Administrative Distance**
  - IBGP: 200
  - EBGP: 20 (preferred over all IGPs)

Some vendors including Cisco also allow EBGP peers to be logically linked over other hops inbetween. This "Multi-Hop" feature might introduce BGP-inconsistency and weakens the reliability as the BGP-TCP sessions cross other routers, so in practice a direct peering should be achieved.

Routing information learned by IBGP messages has much higher administrative distance than information learned by EBGP. Because of this, routes are preferred that do not cross the own autonomous system.

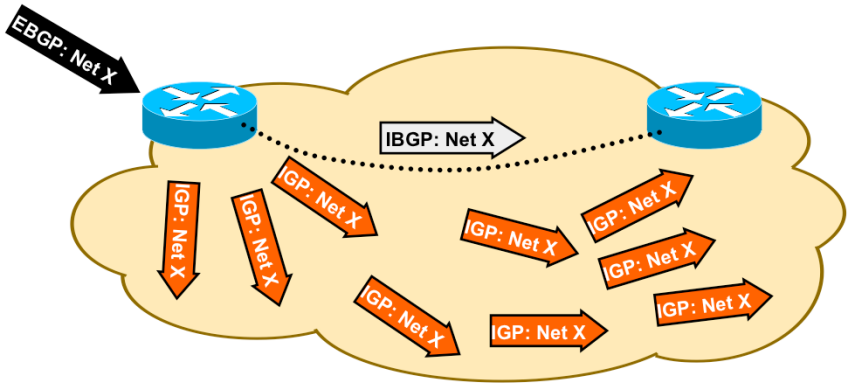
Loop Detection

- Update is only forwarded if own AS number is not already contained in AS_Path
- Thus, routing loops are avoided easily
- But this principle doesn't work with IBGP updates (!)
- Therefore IBGP speaking routers must be fully meshed !!!

For EBGP sessions loop-free topology is guaraneed by checking AS-Path, but it is not the case for IBGP sessions.

BGP ➔ IGP Redistribution

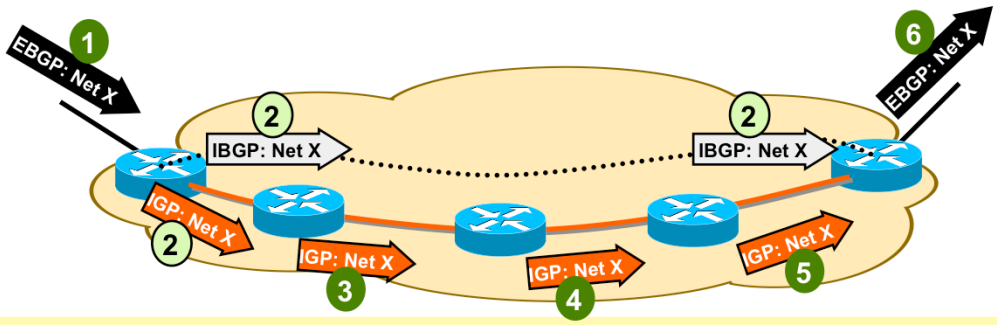
- **Only routes learned via EBGP are redistributed into IGP**
  - To assure optimal load distribution
  - Cisco-IOS default filter behavior



Routes learned via IBGP are never redistributed into IGP. This is the Cisco IOS "default filter" behavior. Obviously, if a router learned a route via IBGP, it is not a external (direct) peer for this route.

Synchronization With IGP

- **Routes learned via IBGP may only be propagated via EBGP if same information has been also learned via IGP**
  - That is, same routes also found in routing table (= are really reachable)
- **Without this "IGP-Synchronization" black holes might occur**



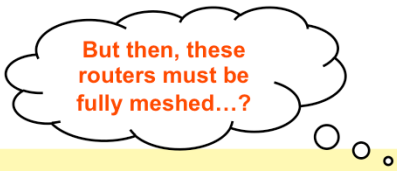
 IP Routing, v6.2 313

When a BGP router learns about an exterior network via an IBGP session, this router does not enter this route into its routing table nor propagates this route via EBGP because the IGP-transit routers might not be aware about this route and therefore convergence has not been occurred yet. The BGP router should propagate the learned route until this route has been entered into its routing table by IGP.

To understand this issue remember that BGP routing information is transported almost instantaneous between two BGP peers, while IGP updates might need quite a long time until reaching the other side of the AS. As illustrated in step 2 in the picture above, the IBGP message has been received by the BGP peer on the right border already, while the first IGP update (advertising the same network X) was injected by the left BGP peer and only reached the next IGP router at this time.

Avoid Synchronization

- **Synchronization with IGP means injecting thousands of routes into IGP**
  - IGP might get overloaded
  - Synchronization dramatically affects BGP's convergence time
- **Alternatives**
  - Set default routes leading to BGP routers (might lead to suboptimal routing)
  - Use only BGP-routers inside the AS !



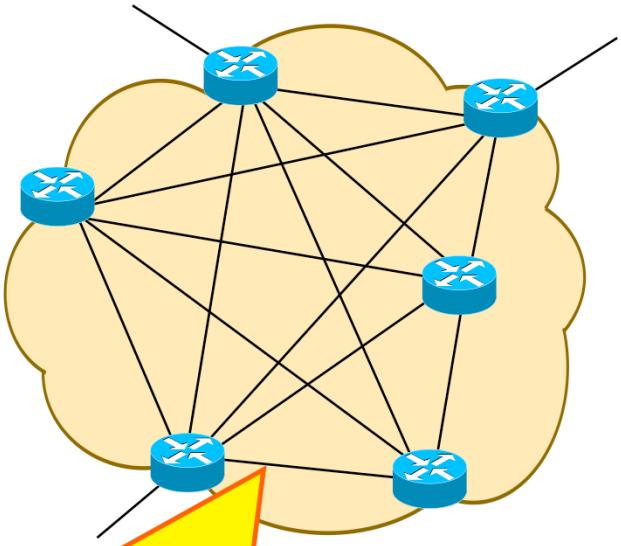
Synchronization is an old idea and leads to unwanted effects. First of all, most IGPs are not designed to carry a huge number of routes as needed in the Internet. Thus IGPs might get overloaded when ten thousands of external routes should be propagated in addition to the interior routes.

Furthermore, external routes are not needed inside an AS and typically a default route pointing to an BGP border router is sufficient (however this might lead to suboptimal routes as the default route might not be the best route). And finally, the consistency of the global BGP routing map would depend on the convergence of several (lots of) IGP routers – a situation that should be avoided!

Note that BGP injection into IGP and required BGP synchronization is not necessary if the AS is a transit AS only, such as many ISP networks. ISP networks have typically BGP routers only and thus need no synchronization. Fortunately many routers today (including Cisco routers) support the option to turn off synchronization.

Fully Meshed IBGP Routers

FYI



- **Does not scale**
  - n(n-1)/2 links
- **Resource and configuration challenge**
- **Solutions:**
  - Route Reflectors
  - Confederations

Note: These are logical IBGP connections! The physical topology might look different!

 IP Routing, v6.2 315

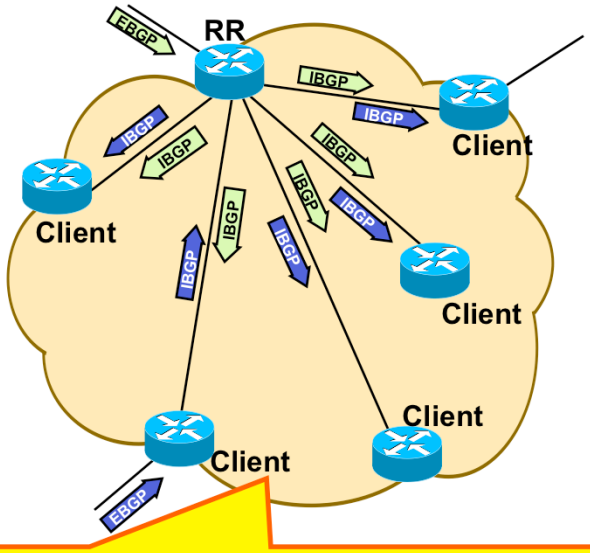
Every BGP router maintains IBGP sessions with all other internal BGP routers of an AS. Obviously, this fully meshed approach does not scale, especially it becomes a resource and manageability problem if the number of BGP sessions in one router exceeds 100.

Remember that each BGP session corresponds to a TCP connection, which requires a lot of system resources. Additionally BGP sessions must be manually established, so a fully meshed environment is also a configuration problem. This is also the reason, why BGP cannot replace traditional IGPs in "normal" autonomous systems. ISPs demand for fast BGP convergence and do not need IGP in general.

Generally, there are two solutions to circumvent this problem: Route Reflectors and Confederations. Both techniques are discussed in the next slides.

Route Reflector

FYI



- RR mirrors BGP messages for "clients"
- RR and clients belong to a "cluster"
- Only RR must be configured
  - Clients are not aware of the RR

Note: Although these are logical IBGP connections, the physical topology should be the main indicator for an efficient cluster design (which router becomes RR)

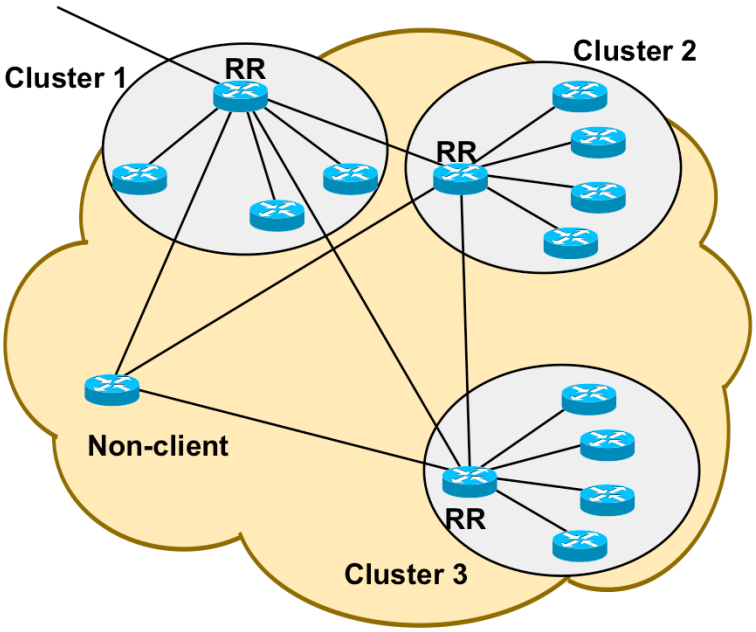
 IP Routing, v6.2 316

Route reflectors are dedicated BGP routers that act like a mirror for IBGP messages. All BGP routers that peer with a RR are called "clients" and belong to a "cluster". Clients are normal BGP routers and have no special configuration – they have no awareness of a RR.

Using RRs there are only n-1 links.

RR Clusters

FYI



- Only RRs are fully meshed
- Special Attributes care for loop-avoidance
- "Non-clients" must be fully meshed with RRs
  - And with other non-clients

Clients are considered as such because the RR lists them as clients.

RR Issues

FYI

- RRs do not change IBGP behavior or attributes
- RRs only propagate best routes
- Special attributes to avoid routing updates reentering the cluster (routing loops)
  - ORIGINATOR_ID

    Contains router-id of the route's originator in the local AS; attached by RR (Optional, Non-Trans.)
  - CLUSTER_LIST

    Sequence of cluster-ids; RR appends own cluster-id when route is sent to non-clients outside the cluster
    (Optional, Non-Transitive)

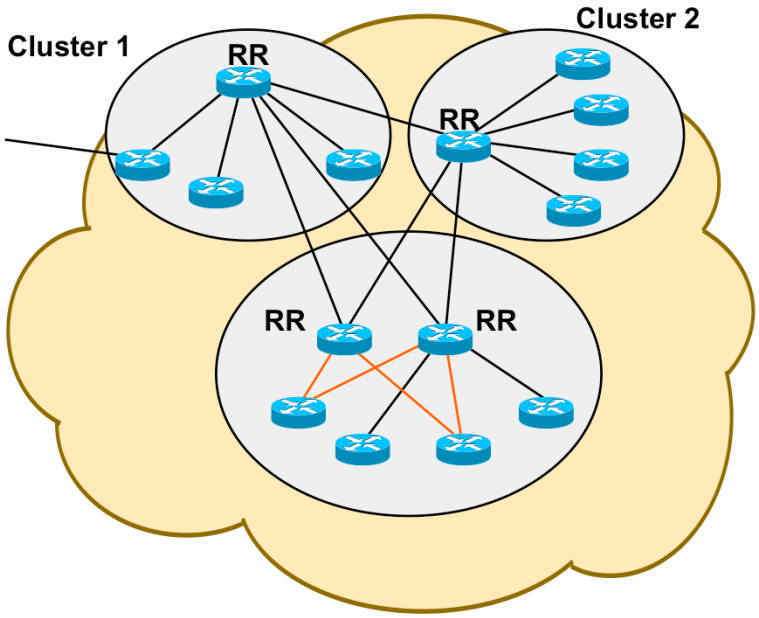
 IP Routing, v6.2 318

It is important to know that RRs preserve IBGP attributes. Even the NEXT_HOP remains the same, otherwise routing loops might occur. Imagine two clusters whose RRs are logically interconnected via IBGP but physically via clients. If one of these RRs learns about a NLRI from the other RR, this RR would reflect that information to its clients – also to that client who forwarded this NLRI information to this RR.

Obviously the NEXT_HOP attribute must remain the same, that is pointing to the RR of the other cluster and not to the local RR, because there is no physical connection between the RRs.

If a RR learns the same NLRI from multiple client peers, only one path will be propagated to other peers. Therefore, when RRs are used, the number of path available to reach a given destination might be lower than that of a fully-meshed approach. Thus, suboptimal routing can only be avoided if the logical topology maps the physical topology as close as possible.

Redundant RRs



- **RR is single point of failure**
  - Other than fully meshed approach
- **Redundant RRs can be configured**
  - Clients attached to several RRs

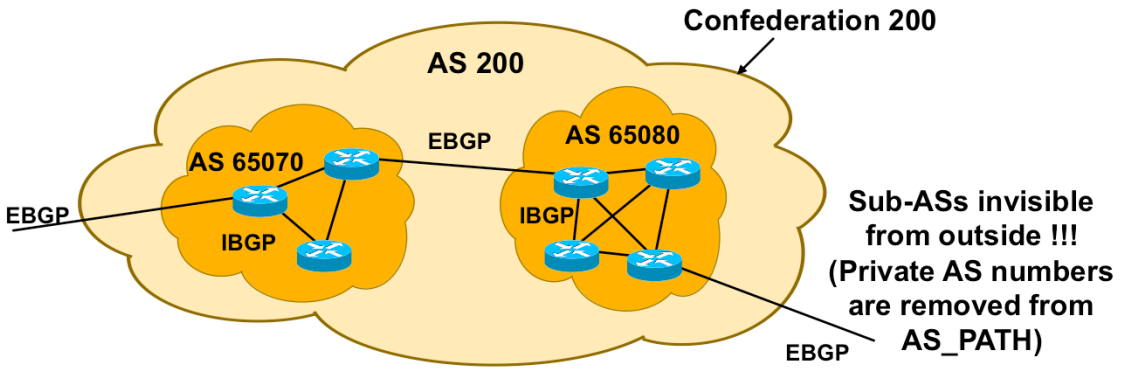
 IP Routing, v6.2 319

Clients are considered as such because the RR lists them as clients.

Confederations

FYI

- Alternative to route reflectors
- Idea: AS can be broken into multiple sub-ASs
- Loop-avoidance based on AS_Path
- All BGP routers inside a sub-AS must be fully meshed
- EBGP is used between sub-ASs



 IP Routing, v6.2 320

Sub-ASs should utilize the private range of AS numbers (64512-65534).

RRs versus Confederations

FYI

- **RRs are more popular**
  - Simple migration (only RRs needs to be configured accordingly)
  - Best scalability
- **Confederations drawbacks**
  - Introducing confederations require complete AS-renumbering inside an AS
  - Major change in logical topology
  - Suboptimal routing (Sub-ASs do not influence external AS_PATH length)
- **Confederations benefits**
  - Can be used with RRs
  - Policies could be applied to route traffic between sub-ASs

Agenda

- **Introduction to IP Routing**
- **RIP**
- **OSPF**
- **Introduction to Internet Routing (BGP, CIDR)**
  - Introduction
  - BGP Basics
  - BGP Attributes
  - BGP Special Topics
  - CIDR **FYI**

Early IP Addressing

- **Before 1981 only class A addresses were used**
  - Original Internet addresses comprised 32 bits (8 bit net-id = 256 networks)
- **In 1981 RFC 790 (IP) was finished and classes were introduced**
  - 7 bit class A networks
  - 14 bits class B networks
  - 21 bits class C networks

 IP Routing, v6.2 323

IP is an old protocol which was born with several design flaws. Of course this happened basically because IP was originally not supposed to run over the whole world.

The classful addressing scheme led to a big waste of the 32 bit address space.

A short address design history:

1980	Classful Addressing	RFC 791	
1985	Subnetting		RFC 950
1987	VLSM		RFC 1009
1993	CIDR		RFC 1517 - 1520

Address Classes

- From 1981-1993 the Internet was Classful (!)
- Early 80s: Jon Postel volunteered to maintain assigned network addresses
  - Paper notebook
- Internet Registry (IR) became part of IANA
- Postel passed his task to SRI International
  - Menlo Park, California
  - Called Network Information Center (NIC)

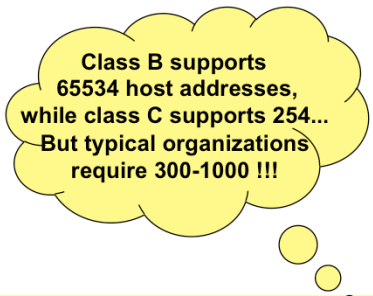
Until 1993 the Internet used classful routing. All organizations were assigned either class A, B, or C network numbers. In the early 1980s, one of the inventors of the Internet, Jon Postel, volunteered to maintain all assigned network addresses—simply using a paper notebook!

Later the Internet Registry (IR) became part of the IANA and Jon Postel's task was passed to the Network Information Center, which is represented by SRI International.

FYI: See http://www.iana.org

Classful – Drawbacks

- **"Three sizes *don't* fit all" !!!**
  - Demand to assign as little as possible
  - Demand for aggregation as many as possible
- **Assigning a whole network number**
  - Reduces routing table size
  - But wastes address space



 IP Routing, v6.2 325

Using the full classes of the addresses it was difficult to match all needs.

Subnetting

- **Subnetting introduced in 1984**
  - Net + Subnet (=another level)
  - RFC 791
  - Initially only statically configured
- **Classes A, B, C still used for global routing !**
  - Destination Net might be subnetted
  - Smaller routing tables

By introduction of subnetting (RFC 791) a network number could be divided into several subnets. Thus large organizations who needed multiple network numbers are assigned a single network number which is further subnetted by themselves. This way, subnetting greatly reduced the Internet routing table sizes and saved the total IP address space.

Routing Table Growth (88-92)

MM/YY	ROUTES ADVERTISED	MM/YY	ROUTES ADVERTISED
Feb-92	4775	Apr-90	1525
Jan-92	4526	Mar-90	1038
Dec-91	4305	Feb-90	997
Nov-91	3751	Jan-90	927
Oct-91	3556	Dec-89	897
Sep-91	3389	Nov-89	837
Aug-91	3258	Oct-89	809
Jul-91	3086	Sep-89	745
Jun-91	2982	Aug-89	650
May-91	2763	Jul-89	603
Apr-91	2622	Jun-89	564
Mar-91	2501	May-89	516
Feb-91	2417	Apr-89	467
Jan-91	2338	Mar-89	410
Dec-90	2190	Feb-89	384
Nov-90	2125	Jan-89	346
Oct-90	2063	Dec-88	334
Sep-90	1988	Nov-88	313
Aug-90	1894	Oct-88	291
Jul-90	1727	Sep-88	244
Jun-90	1639	Aug-88	217
May-90	1580	Jul-88	173

Growth in routing table size, total numbers
Source for the routing table size data is MERIT

The list above shows the growth of the routing tables from 1988 until 1992 in total numbers.

Network Number Statistics, April 1992

	Total	Allocated	Allocated %
Class A	126	48	54%
Class B	16383	7006	43%
Class C	2097151	40724	2%

Only 2% of more than 2 million Class C addresses assigned !!!

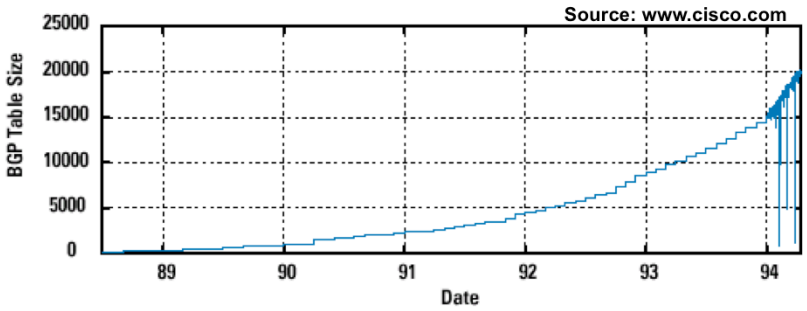
Source: RFC 1335

 IP Routing, v6.2 328

The table above shows a statistic for the assignment of IP addresses in April 1992. Obviously, class A and B addresses have been allocated quicker than class C addresses. In the following years the utilization of class C addresses increased rapidly while class A and B addresses were spared.

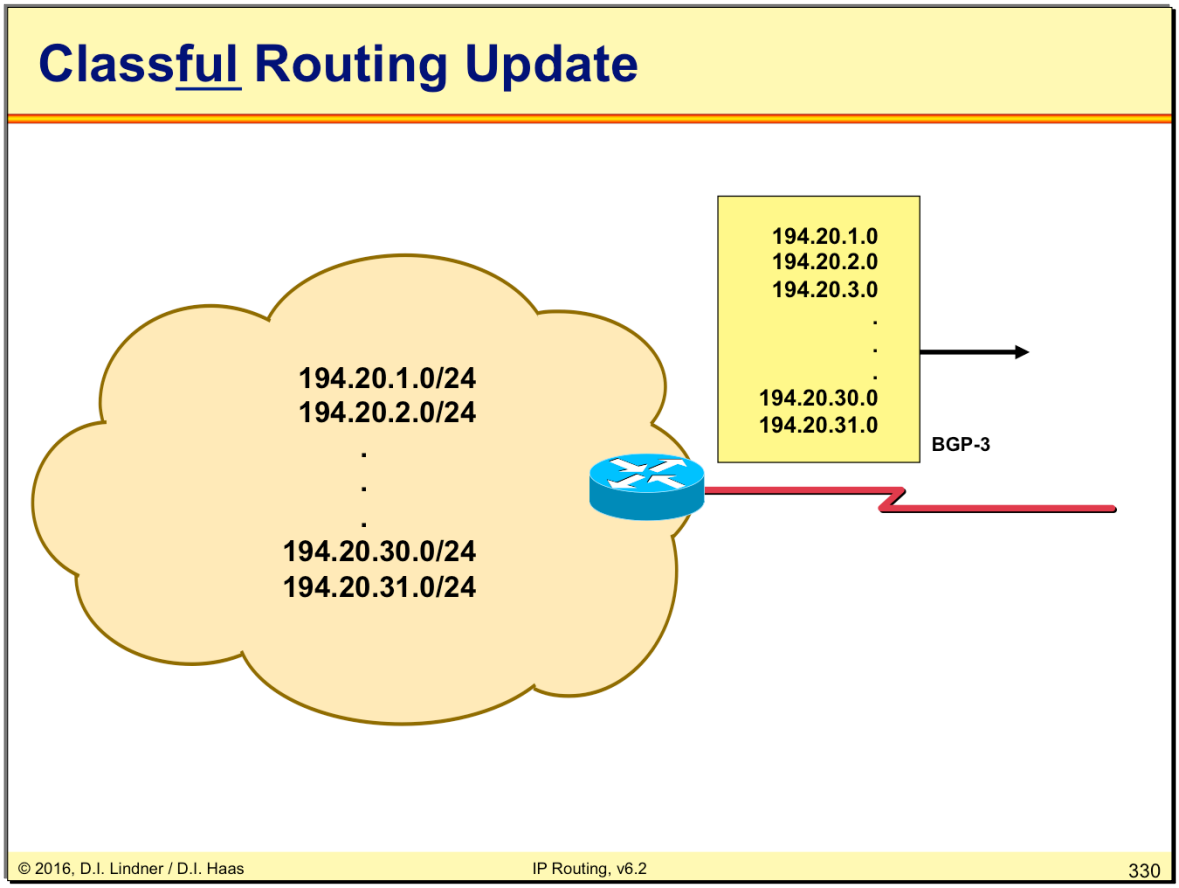
Especially VLSM and NAT (invented 1994) supported the utilization of class C addresses.

Supernetting (RFC 1338)

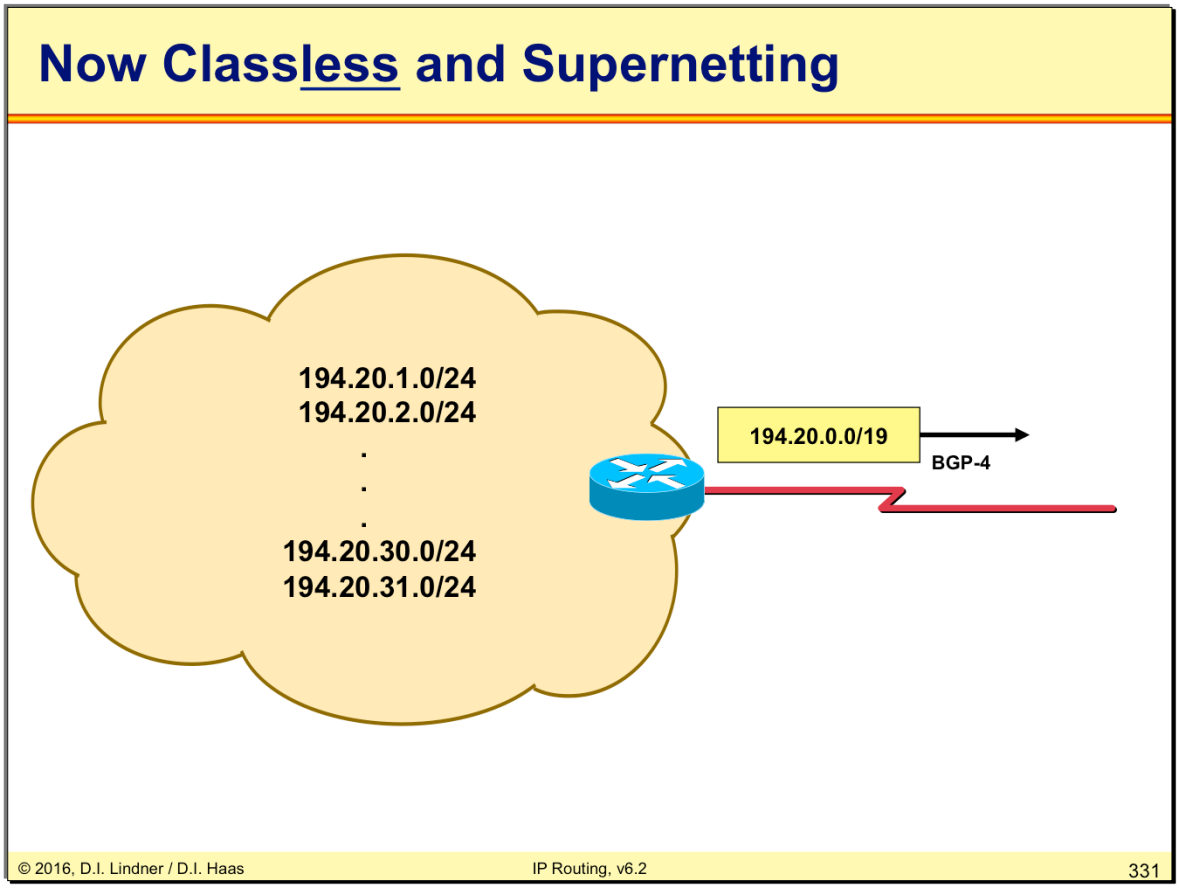


- **In 1992: RFC 1338 stated scaling problem:**
  - Class B exhaustion
  - No class for typical organizations available
  - Unbearable growth of routing table
- **Use subnetting technique also in the Internet !**
  - Do hierarchical IP address assignment !
  - Aggregation = "Supernetting"
    (Smaller netmask than natural netmask)

RFC 1338 introduced Supernetting: an Address Assignment and Aggregation Strategy, now obsoleted by RFC 1519.



BGP-3 was a classful routing protocol, sending the information about major class A, B, and C networks only.



BGP-4 is classless, it can aggregate a range of class C network in one supernet.

CIDR

- **September 1993, RFC 1519: Classless Inter-Domain Routing (CIDR)**
- **Requires classless routing protocols**
  - BGP-3 upgraded to BGP-4
  - New BGP-4 capabilities were drawn on a napkin, with all implementers of significant routing protocols present (legend)
  - RFC 1654

RFC 1519 introduced Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy

RFC 1654 a draft standard for BGP – 4

RFC 1771 a standard for BGP - 4

Address Management

- ISPs assign
  contiguous blocks of
  contiguous blocks of
  contiguous blocks ...
  of addresses to their customers
- Aggregation at borders possible !
- Tier I providers filter routes with prefix lengths larger than /19
  - But more and more exceptions today...

To minimize the sizes of the routing tables ISPs use agregation, giving the customers the contiguous blocks of networks or subnets. Most of the ISPs would not accept routes from other ISP if the prefix is longer than /19.

International Address Assignment

- **August 1990, RFC 1174 (by IAB) proposed regionally distributed registry model**
  - Regionally means continental ;-)
- **Regional Internet Registries (RIRs)**
  - RIPE NCC
  - APNIC
  - ARIN

RFC 1174 IAB Recommended Policy on Distributing Internet Identifier Assignment.

This RFC represents the official view of the Internet Activities Board (IAB), and describes the recommended policies and procedures on distributing Internet identifier assignments and dropping the connected status requirement.

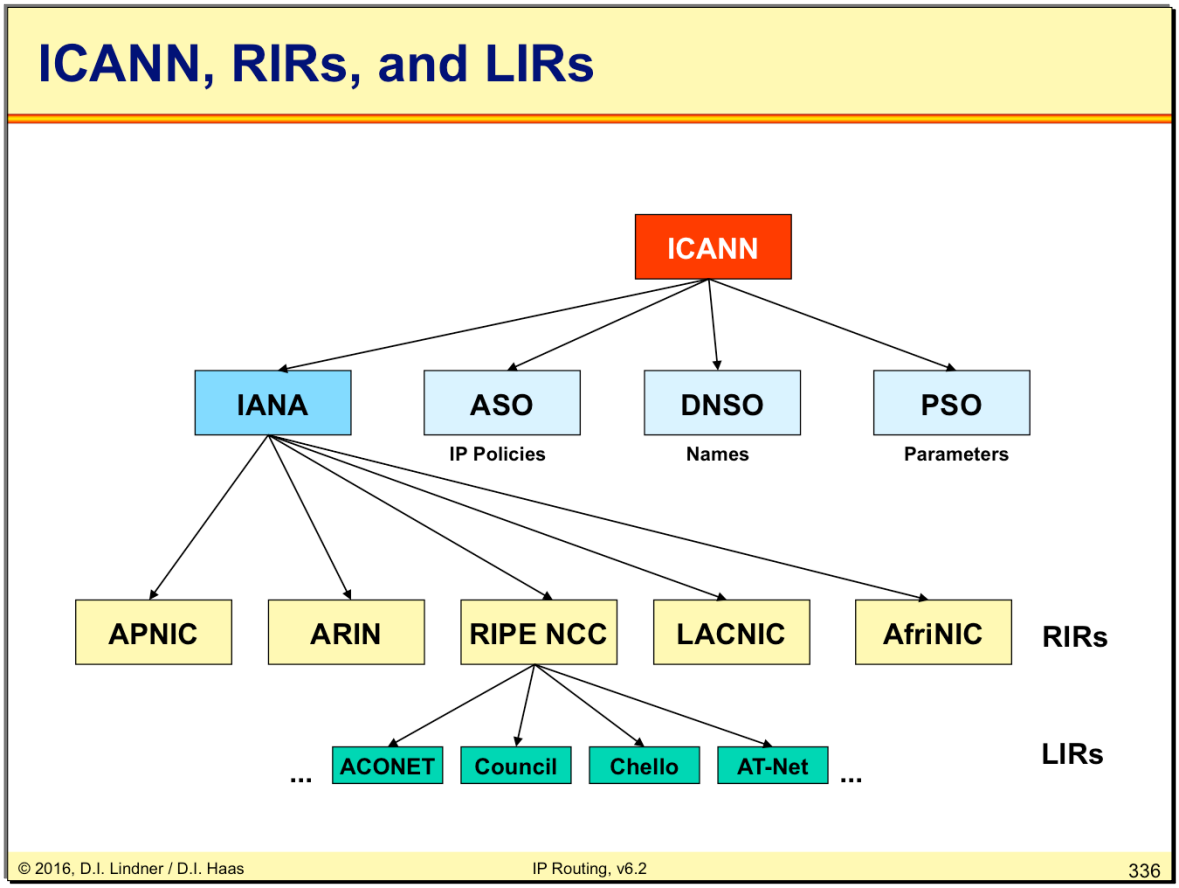
RIRs

- **RIPE NCC (1992)**
  - Réseaux IP Européens (RIPE) founded the Network Coordination Centre (NCC)
- **APNIC (1993)**
  - Asia Pacific Information Centre
- **ARIN (1997)**
  - American Registry for Internet Numbers
- **AfriNIC**
  - Africa
- **LACNIC**
  - Latin America and Caribbean

RIPE NCC is located in Amsterdam and serves 109 countries including Europe, Middle-East, Central Asia, and African countries located north of the equator. The RIPE NCC currently consists of more than 2700 members.

APNIC was relocated to Brisbane (Australia) in 1998. Currently there are 700 member organizations. Within the APNIC there are also five National Internet Registries (NIRs) in Japan, China, Korea, Indonesia, and Taiwan, representing more than 500 additional organizations.

AfriNIC and LACNIC are relatively new RIRs (2002).



After foundation of the ICANN, the Internet Assignment Numbers Authority (IANA) is only responsible for IP address allocation to RIRs.

Other sub-organizations of the ICANN:

- Address Supporting Organization (ASO), which was founded by APNIC, ARIN, and RIPE NCC, and should oversee the recommendations of IP policies
- Domain Name Supporting Organization (DNSO) is responsible for maintaining the DNS
- Protocol Supporting Organization (PSO) is responsible for registration of various protocol numbers and parameters used by RFC protocols

Originally, all tasks of these sub-organizations were performed by the IANA only. Today the IANA only cares for address assignment to the RIRs.

The slide above shows a few of the long list of LIRs in Austria. These LIRs are those who are widely known by Internet users as "Internet Service Providers".

CIDR Concepts Summary

- Coordinated address allocation
- Classless routing
- Supernetting

RFC 1366 Address Blocks

- 192.0.0.0 - 193.255.255.255 ... Multiregional
- 194.0.0.0 - 195.255.255.255 ... Europe
- 198.0.0.0 - 199.255.255.255 ... North America
- 200.0.0.0 - 201.255.255.255 ... Central/South America
- 202.0.0.0 - 203.255.255.255 ... Pacific Rim

RFC 1366 Guidelines for Management of IP Address Space, was obsoleted by 1466 in 1993, in 1996 an RFC 2050 came out.

Class A Assignment

- **IANA responsibility**
  - RFC 1366 states: *"There are only approximately 77 Class A network numbers which are unassigned, and these 77 network numbers represent about 30% of the total network number space."*
- **64.0.0.0 – 127.0.0.0 were reserved for the end of (IPv4) days ?**
  - Recent assignments (check IANA website)

The Class A addresses assignment is controlled by the IANA.

Class B Assignment

- **IANA and RIRs requirements**
  - Subnetting plan which documents more than 32 subnets within its organizational network
  - More than 4096 hosts
- **RFC 1366 recommends to use multiple Class Cs wherever possible**

In order to receive a class B address, an organization must fulfill strict requirements such as employing more than 4096 hosts and more than 32 subnets.

Class C Assignment

- If an organization requires more than a single Class C, it will be assigned a bit-wise contiguous block from the Class C space
- Up to 16 contiguous Class C networks per subscriber (= one prefix, 12 bit length)

Organization	Assignment
1) requires fewer than 256 addresses	1 class C network
2) requires fewer than 512 addresses	2 contiguous class C networks
3) requires fewer than 1024 addresses	4 contiguous class C networks
4) requires fewer than 2048 addresses	8 contiguous class C networks
5) requires fewer than 4096 addresses	16 contiguous class C networks

Example (RFC 1366) for Class C assignment:

For instance, an European organization which requires fewer than 2048 unique IP addresses and more than 1024 would be assigned 8 contiguous class C network numbers from the number space reserved for European networks, 194.0.0.0 - 195.255.255.255. If an organization from Central America required fewer than 512 unique IP addresses and more than 256, it would receive 2 contiguous class C network numbers from the number space reserved for Central/South American networks, 200.0.0.0 - 01.255.255.255.

RFC 1918 – Private Addresses

- **In order to prevent address space depletion, RFC 1918 defined three private address blocks**
  - 10.0.0.0 - 10.255.255.255 (prefix: 10/8)
  - 172.16.0.0 - 172.31.255.255 (prefix: 172.16/12)
  - 192.168.0.0 - 192.168.255.255 (prefix: 192.168/16)
- **Connectivity to global space via Network Address Translation (NAT)**

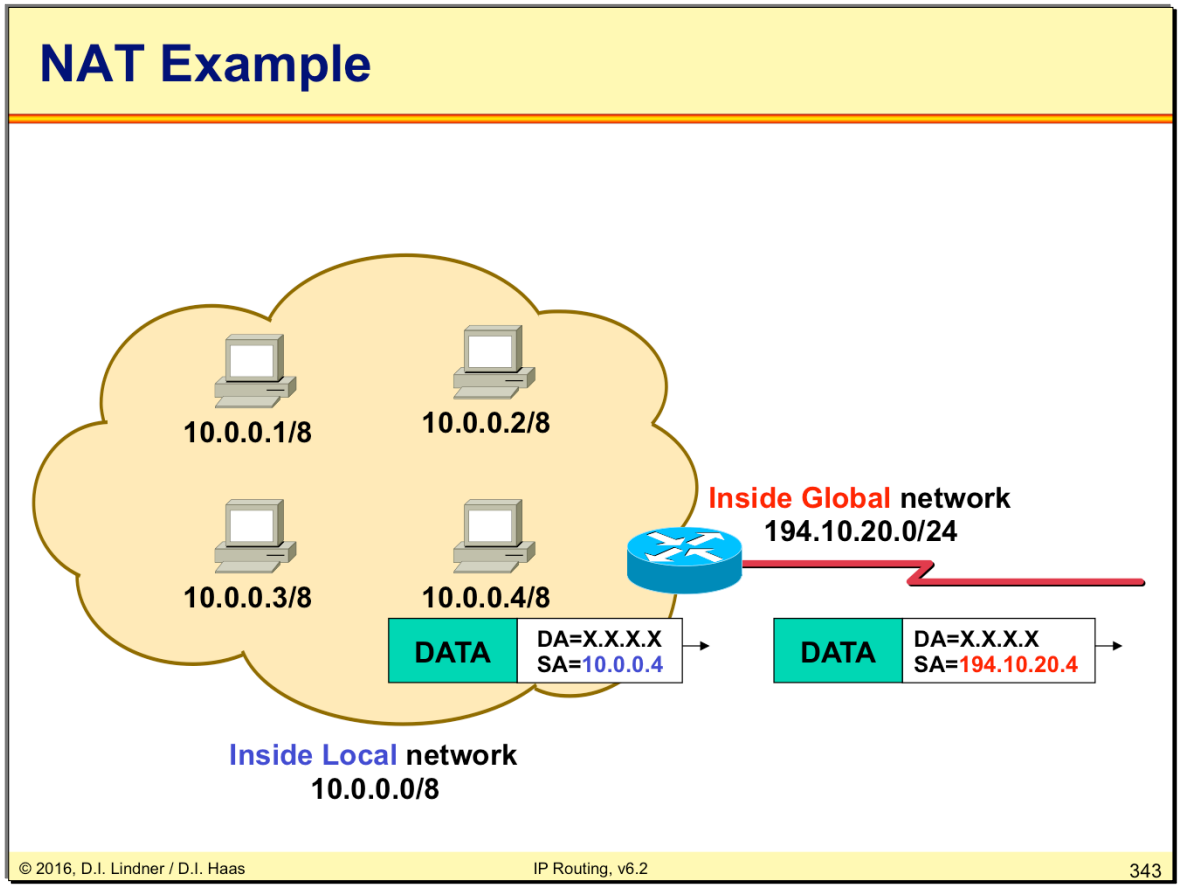
 IP Routing, v6.2 342

RFC 1918 defines an "Address Allocation for Private Internets", that is three address spaces, which should only be used in private networks.

Any route to this network must be filtered in the Internet! Any router in the Internet must not keep any RFC 1918 address in its routing table!

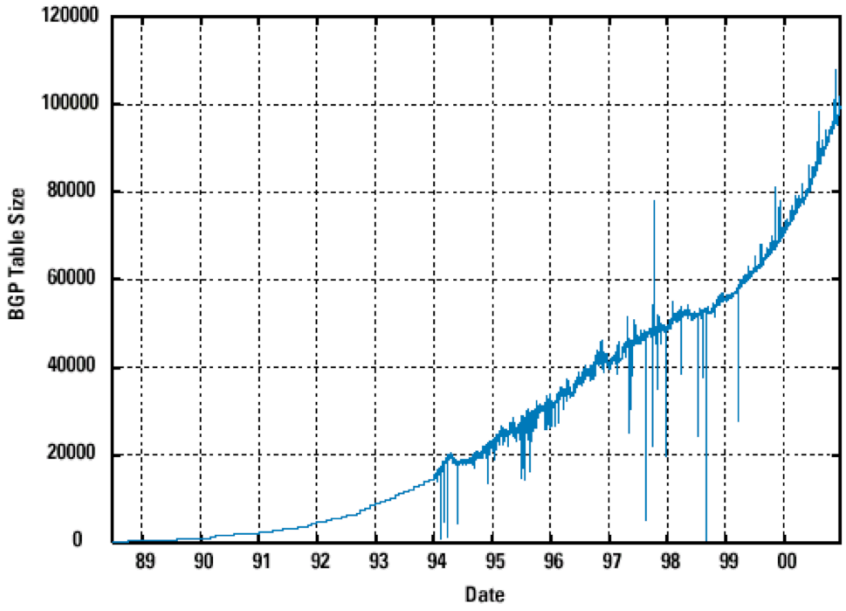
Together with these addresses, Network Address Translation (NAT) is needed if private networks should be connected to the Internet.

This solution greatly reduces the number of allocated IP addresses and also the routing table size because now class C networks can be assigned very efficiently, using a prefix up to /30.



Network Address Translation (NAT) In order to be able to communicate with Internet we have to translate private addresses (inside local) into official, assigned by an ISP (inside global).

But...



Source: www.cisco.com

 IP Routing, v6.2 344

But this is not really the end of the story. The growth rate of the Internet was and is generally exponential, that is exp(k*x). Soon after the introduction of CIDR the progressive factor k increased dramatically, thus even CIDR could only reduce k, but not the general exponential character.

It is interesting to question how long the (also exponential) growth rate of silicon memory and processing power together with CIDR and NAT can mitigate the effects of the Internet growth.

As for today, the only solution to deal with this problem in the long run is to introduce IPv6 and a more hierarchical routing strategy.